Manufacturing at double the speed

Julian M. Allwood\textsuperscript{a}, Vadim A. Baines-Jones\textsuperscript{b}, Tom H.C. Childs\textsuperscript{c}, Adam T. Clare\textsuperscript{d}, Anjali K.M. De Silva\textsuperscript{e}, Vimal Dhokia\textsuperscript{f}, Ian M. Hutchings\textsuperscript{g}, Richard K. Leach\textsuperscript{d}, Wayne Leahy\textsuperscript{h}, David R. Leal-Ayala\textsuperscript{i}, Stewart Lowth\textsuperscript{d}, Candice E. Majewski\textsuperscript{b}, Adelaide Marzano\textsuperscript{l}, Jörn Mehnen\textsuperscript{j}, Aydin Nassehi\textsuperscript{l}, Erdem Ozturk\textsuperscript{k}, Mark H. Raffles\textsuperscript{g}, Raj Roy\textsuperscript{l}, Islam Shyha\textsuperscript{j}, Sam Turner\textsuperscript{k}.

\textsuperscript{a}Department of Engineering, University of Cambridge

\textsuperscript{b}Fives Landis Ltd.

\textsuperscript{c}School of Mechanical Engineering, University of Leeds

\textsuperscript{d}Institute for Advanced Manufacturing, University of Nottingham

\textsuperscript{e}Glasgow Caledonian University

\textsuperscript{f}Department of Mechanical Engineering, University of Bath

\textsuperscript{g}Element Six Ltd.

\textsuperscript{h}Department of Mechanical Engineering, University of Sheffield

\textsuperscript{i}School of Mechanical and Aerospace Engineering, Queen's University Belfast

\textsuperscript{j}Manufacturing Department, Cranfield University

\textsuperscript{k}Advanced Manufacturing Research Centre, University of Sheffield

\textsuperscript{l}Department of Mechanical & Construction Engineering, Northumbria University
Abstract

The speed of manufacturing processes today depends on a trade-off between the physical processes of production, the wider system that allows these processes to operate, and the co-ordination of a supply chain, innovation, management and the financial system. Could the speed of this activity be doubled? This paper begins to address this hypothetical question, starting with examination of a diverse set of case studies spanning the activities of manufacturing. This reveals that the constraints on increasing manufacturing speed have some common themes, and several of these are explored in more detail, to identify absolute limits to performance. The physical processes of production are constrained by factors such as machine stiffness, actuator acceleration, heat transfer and the delivery of fluids, and for each of these, a simplified model is used to analyse the gap between current and limiting performance. The wider systems of production require the co-ordination of resources and push at the limits of human biophysical and cognitive limits, and evidence about these is explored and related to current practice. Out of this discussion, five promising innovations are explored to show examples of how manufacturing speed is increasing – with line arrays of point actuators, parallel tools, tailored application of precision, hybridisation and task taxonomies. The paper addresses a broad question which could be pursued by a wider community and in greater depth, but even this first examination suggests the possibility of unanticipated innovations in current manufacturing practices.

**Keywords:** Manufacturing; Speed; Productivity; Constraints
1. Introduction

Research in manufacturing often develops incrementally, typically by extending known methods of analysis and applying them to known processes, or by extending existing processes or systems. But does this incremental approach fail to identify opportunities for larger step-changes in processes, systems or analysis which may arise when familiar methods are challenged? This paper aims to stimulate a search for such opportunities, by posing an artificial question: is manufacturing at double the speed possible?

The paper arises out of a meeting organised by CIRP UK in May 2014. Sixteen researchers responded to the challenge “is manufacturing at double the speed possible?” based on the experience and knowledge they have accrued in their own activities. Following the talks, around 40 meeting attendees worked to develop a structure within which an answer to the question could be organised. After the meeting, the presenters wrote up their own contributions within this structure, and out of their input, this paper was created.

The intention of examining manufacturing at double the speed is to develop an improved understanding of the limits to current processes. The paper therefore assumes that for each process discussed, the workpiece material must be as specified by the designer: in practice, material selection could be iterative and manufacturing speed might be increased by substitution of a different component material, but this possibility is excluded. The scope of the paper further excludes two approaches to increasing manufacturing speed which are already normal practice in industry:

- The speed of manufacturing for many components or products could be increased by increasing the capacity or specification of the process – for example by running machines in parallel, increasing machine power, or switching between existing processes. Similarly, speed could be increased by reducing requirements for tolerance or quality.

- For many components or products produced in small volumes, or over only a brief lifespan, the speed of manufacturing will increase with experience, as commonly characterised by a ‘learning curve.’ The
steady improvement of performance achieved by experiments leading to better tool path design, tool or lubricant selection, or vibration control for example, is already normal practice.

This regular process of improvement is, in normal industrial practice, often addressed with a trial and error approach, with manufacturing improvements made incrementally until business pressures dictate that efforts are focused elsewhere. Formal approaches to process improvement include the use of Design of Experiments, critical path analysis and the use of process simulation. Ozturk et al. (2009) contrast improvement processes with and without simulation, and point out that process simulation tools generally require expert users so are more common in research institutes and with technology providers, and are less widely used and not integrated into manufacturing systems in practice.

In contrast to this regular practice of improvement, the scope of the paper is to explore whether the speed of today's manufacturing practices can be doubled. The answer to this question must acknowledge that speed limits can arise from any of the components of the “rotating bottlenecks” in fig. 1.

![Figure 1. Rotating bottlenecks that inhibit manufacturing at double the speed](image)

The figure illustrates that limits to manufacturing speed may arise from:

- the materials process itself – the response of the workpiece to the boundary conditions created by the process. For example, heat treatment may be limited by the size and thermal conductivity of the workpiece.
the system – the materials process is operated by people with constrained capabilities, and for example, product quality can only be improved up to the performance level of the metrology used.

the co-ordination of the process with the rest of the system – the process can only operate when loaded with a workpiece in the right state, and appropriately set up and programmed. For example, materials handling, CNC programming and equipment testing may limit the speed of switchover between parts in a job shop.

In industrial practice, the limits to manufacturing speed may arise from any of these three bottlenecks, and improvement in one area may simply reveal a different and equally constraining bottleneck elsewhere. However, the interactions in fig. 1 are context specific, so the paper aims to examine doubling the speed of manufacturing by overcoming each form of bottleneck separately.

Section 2 of the paper presents a survey of the limits to performance of a set of exemplar manufacturing activities, those most familiar to the authors. These examples are used to structure the exploration of limits to manufacturing speed in sections 3 and 4, considering each of the three bottlenecks of fig. 1 in turn. Section 5 presents several promising examples of significant speed increases leading to a brief discussion of the implications of the paper in section 6.

2. Barriers to doubling the speed of manufacturing today

What would go wrong if the speed of manufacturing was doubled today? This section presents a series of case studies of current manufacturing activities to explore this question. Figure 2 shows how these case studies relate to the three bottlenecks of fig. 1, with the materials processing operations sub-classified according to whether they add to, maintain (for example by changing shape or by heat treatment) or subtract from the mass of the workpiece.
Figure 2. Case studies of current manufacturing activities and their speed limits

The five sub-sections that follow discuss the constraints on doubling speed in each of the case study groups in fig. 2 in turn. The case studies were selected according to the authors’ research interests, so are a representative but not exhaustive sample of today’s manufacturing processes, but in many cases the causes of their speed limits are related. For example, both strip rolling and machining can lead to equipment vibration: the materials process driving the vibration is different in the two cases, but the options to modify the equipment to suppress the vibration are similar.

2.1 Material processing that subtracts mass from a component

The processes discussed in this section all reduce the mass of a component, and are organised according to the form of material removed – whether as chips, powder, reaction products or vapour. The processes are mainly applied to metals, but composites and polymers are also considered.

2.1.1 Machining

In machining processes, see fig. 3, a tool removes chips of material from a workpiece at a rate that depends on the depth of cut and the speed at which the tool moves relative to the workpiece. The operation of machining processes aims
to maximise tool speed while maintaining part quality and minimising damage to the tool. In turn:

- the maximum speed of cutting is limited by vibration and the acceleration capability of the machine;

- part quality is defined by the geometric accuracy of the surface including burrs or tear-outs, and the degree of damage created in the sub-surface (where strain hardening and residual stress lead to void growth), both of which are influenced by the control of heat flows around the tool;

- tools are damaged in use by wear which changes the geometry of the tip, in turn influencing heat generation and tool forces, which will eventually lead to tool breakage.

**Figure 3. Summary of the limits to machining speed.**

Vibration in machining, according to Altintas and Weck (2004), may be forced – i.e. related to the speed of tool motion – in which case it is relatively easy to control, or self-excited – related to the stiffness of the machining system – known as 'chatter'. A typical representation of the stability of a system to chatter vibrations is shown in fig. 4, with the margin of instability characterised by 'lobes' within which larger depths of cut are possible. Increasing cutting speed increases the size of stability lobes and increasing the stiffness of the system increases the magnitude of them even more, but this stiffness is a complex function of the current configuration of the machine, workpiece and fixtures and it may vary
throughout an operation. Prediction of the stability diagram is therefore a current area of research (for example Budak and Altintas (1998) on frequency domain, Sims (2005) on time-domain, or Insperger et al. (2003) on semi-discretisation models). In parallel, research is exploring the development of passive vibration control (for example shunt damping (Erturk and Inman, 2008), tuned mass damper systems (Tarng et al., 2000), or airbag dampers (Geng, 2008)) or active vibration control (such as electromagnetic guides (Denkena et al., 2014), inertial actuators (Munoa et al., 2013) or active piezoelectric damping (Zhang and Sims, 2005)).

![Stability Diagram]

**Figure 4.** Representative stability diagram showing stable and unstable regions

Tool speed in turning and drilling operations is a function of motor power and essentially unlimited, but in milling, where the tool follows a complex path, tool speed is limited by acceleration. Experience of machine design suggests that the acceleration of an axis is often constrained by the moment of inertia of the motor rotor itself (because high ratio gearboxes are used to convert motor rotation into axis motion) but rapid rates of change in acceleration may also lead to an impulsive force on the machine that triggers vibration at machine resonant frequencies, and this is now addressed by ‘jerk’ control.

Surface quality in machined and turned components is principally defined by geometry which is a function of the final machining step only, and determined by machine stiffness, feed rate (slow feeds are used to limit vibration and heat
generation), bearing precision and predictable tool geometry. However, for some workpiece materials, surface quality may also be constrained by the material – for example, second phase particles of titanium-aluminium alloys may be brittle and crack, limiting surface roughness to the size of the particles. In drilling operations, surface quality also relates to damage at the exit of the hole from the workpiece. This may include exit burrs when drilling metals, or delamination, fuzzing and spalling when drilling composites (Shyha et al., 2009), and can be countered by use of pilot holes, back-up plates (Hocheng and Tsao, 2005) or reduced feed rates near to the exit (Khashaba, 2013). Sub-surface quality defects, typically in the depth range 10’s to 100’s of micro-meters beneath the surface, arise when plastic deformation beneath the tool or heat generated by its passing leads to (tensile) residual stress, strain hardening and the accumulation of voids which eventually become micro-cracks.

Some recent evidence shown in fig. 5 demonstrates that doubling tool hardness may lead to a threefold increase in the number of components machined prior to tool failure. With increasing wear, tool breakage becomes more likely and as this can be expensive either through damage of a nearly complete component after extensive processing or, as Liang et al. (2004) report, through tool failure causing ~20% of machine tool downtime, it is managed through early tool replacement based on statistical models of tool life (Childs et al., 2000). Shyha et al. (2011) also report that in drilling, chip clogging of the curled and deformed swarf within drill flutes may cause premature tool failure which can be countered by greater cutting fluid pressure or tool geometry modifications.
Manufacturers today will often choose to use tools at lower speeds to extend their life, in order to minimise overall costs. In a single machining operation, tool life time $T$ reduces more rapidly than the inverse of increases in cutting speed $V$. In general $VT^n = C$ with $C$ and $n$ constants depending on the tool / work combination and $0.1 < n < 0.8$. As speed is increased a tool will cut less material before it must be changed. Eventually the return from increased speed is outweighed by the loss from increased downtime and tool replacement costs. To examine this trade off, equations 1 predict the cutting speed $v_p$ for maximum productivity and the ratio of $v_p$ to the speed for minimum cost, $v_m$.

$$v_p = \left( \frac{C}{t_{ct}} \right) \left( \frac{n}{(1-n)} \right)^n$$

$$\frac{v_p}{v_m} = \left( 1 + \frac{C_t}{Mt_{ct}} \right)^n$$

Figure 5. Evidence of the potential for tool life increase through use of harder tool materials: (a) Tool life in turning of metal matrix composite components related to tool hardness (after Cooper et al., 2005); (b) tool life when milling silicon aluminium with tools of three materials.
They are based on work from more than 100 years ago by Taylor (1906) and apply to speed selection for a single machining operation. $C_t$ is the cost of a replaced tool, $M$ the time cost (machine, worker and overhead cost), $t_{ct}$ the time for replacement. With typical values of the parameters, Childs et al. (2000a) report that $v_p/v_m$ can certainly be of the order 2.0, although Childs et al. (2000b) observe that tool life shows statistical variance (Childs et al., 2000b) so could be represented by a range of $C$ values. The consequence of (1) is that manufacturers who choose to machine at minimum cost could often double the speed of their operation – but this would increase costs.

Heat is generated in machining by the plastic work of chip formation and friction between the flank face of the tool and the workpiece, and the rake face of the tool and the chip. Heat is removed as a consequence of chip formation, by diffusion into the workpiece and tool, and through the action of cutting fluids, and the balance of heat addition and removal is referred to by Malkin and Guo (2007) as a problem of thermal management. This management of heat is discussed in detail in the section on grinding below, but a particular challenge in machining operations is to apply cutting fluids to the interface between tool and workpiece: except at low speeds, the tool tends to push the fluid away and the continuous creation of new surface as part of chip formation mitigates against the delivery of fluid to the point of maximum heat generation. Concerns over health and safety and cost are currently motivating the pursuit of machining without lubricants, although cooling processes remain important in precision machines where the temperature of every component in the machine must be regulated to within $\pm 5^\circ\text{C}$ (in an example by Bryan, 1990) to avoid thermal distortions that limit product precision.

The direction of sliding of a chip over a tool during metal machining keeps fluid lubricant from the chip/tool contact. The contact shear stress is typically the shear flow stress $k$ of the chip material (except for free-machining alloys with solid lubricant inclusions). The maximum tool temperature $T_{\text{max}}$ caused by sliding increases as the cutting speed $v_c$ and uncut chip thickness $h$ are increased. Based on models and data in Childs et al. (2000c) eqn. 2 and Table 1 provide
relations between work and tool material properties and recommended values of $h_{VC}$ and $T_{max}$.

$$T_{max} \approx \frac{k}{\rho C} \left( 2 + 0.5 \beta \left( \frac{h_{VC}}{\kappa} \right)^{0.5} \right)$$

Equation (2) relates $T_{max}$ to $k$ and $h_{VC}$, with $\rho C$ and $\kappa$ the heat capacity and thermal diffusivity of the chip material, and $\beta$ depending on the ratio of tool to chip thermal conductivity. $\beta \approx 1.0$ except for Ti alloys for which it can fall to 0.5.

Equation (2) can be inverted to provide a value of $h_{VC}$ that generates a particular $T_{max}$.

<table>
<thead>
<tr>
<th>Chip material</th>
<th>$k/\rho C$ (°C)</th>
<th>$\kappa$ (mm$^2$/s)</th>
<th>$T_m$ or $T_d$ (°C)</th>
<th>$h_{VC_{max}}$ (mm$^2$/s)</th>
<th>Recommended $h_{VC}$ (mm$^2$/s) and $T_{max}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cu alloys</td>
<td>110</td>
<td>25</td>
<td>1000</td>
<td>5000</td>
<td>Finishing 500 (460°C), Roughing 2000 (700°C)</td>
</tr>
<tr>
<td>Al alloys</td>
<td>100</td>
<td>70</td>
<td>500</td>
<td>2500</td>
<td>Finishing 600 (350°C), Roughing 1500 (430°C)</td>
</tr>
<tr>
<td>Carbon steels</td>
<td>150</td>
<td>10</td>
<td>1200</td>
<td>1400</td>
<td>Finishing 250 (670°C), Roughing 1000 (1050°C)</td>
</tr>
<tr>
<td>Ni-Cr alloys</td>
<td>150</td>
<td>4</td>
<td>1200</td>
<td>600</td>
<td>Finishing 60 (590°C), Roughing 120 (710°C)</td>
</tr>
<tr>
<td>Ti alloys</td>
<td>270</td>
<td>4</td>
<td>1200</td>
<td>380</td>
<td>Finishing 75 (830°C), Roughing 200 (1020°C)</td>
</tr>
</tbody>
</table>

Table 1. Material properties and maximum $h_{VC}$; also recommended $h_{VC}$ (Sumitomo, 2014) and associated $T_{max}$. Material properties and recommended $h_{VC}$ are mid-range values.

The left columns of Table 1 apply the inversion to estimating $h_{VC_{max}}$ from values of $k/(\rho C)$ and $\kappa$ and maximum allowable values of $T_{max}$, either $T_m$ or $T_d$. In the case of Cu and Al alloys $T_{max}$ is the melting temperature of the chip ($T_m \approx 1000$ or $500$°C respectively). In all other cases it is degradation of the tool. For example, cobalt binder reacts with the carbides of cemented carbides ($T_d = 950$ to 1150°C), inter-granular phases of ceramic tools liquefy ($T_d = 1350$ to 1500°C), cubic boron nitride reverts to hexagonal form ($T_d = 1100$ to 1350°C). In Table 1 $T_d$ is taken to be 1200°C for all such cases, for simplicity. The right two columns of Table 1 are recommended values of $h_{VC}$ for finishing and roughing, from a commercial tool catalogue (Sumitomo, 2014). Temperatures estimated from these, from eqn. (2), are included, in brackets. The generalizations of Table 1 hide many special cases, but reflects general experience of increasingly difficult machinability (reducing allowable $h_{VC_{max}}$) from Cu to Al to Fe to Ni-Cr to Ti alloys.

Table 1’s values of $T_{max}$ in finishing conditions, typically 0.5 to 0.7 $T_m$ or $T_d$, indicate some scope for increasing manufacturing speed. The challenge is to develop tool materials that show less wear than currently, in un-lubricated
conditions that generate contact temperatures from 600 to 800 °C (for Fe, Ni-Cr and Ti alloys). Such wear will be thermally activated, diffusive or chemical in origin. For the popular ceramic coated cemented carbide tools, it will also include thermal fatigue mechanisms arising from the mismatch in thermal expansion between the coating and the substrate.

The physical limits to machining speed revealed in this discussion may be grouped under three headings: tool thermal collapse (in roughing conditions), tool wear, surface integrity and residual stress (in finishing conditions) and tool acceleration and machine stiffness. Which is the governing limit, and how close to the limit is existing practice, depends on what is the work material and machining operation. Table 1 indicates that recommended material removal rates for roughing are mostly within a factor two of the tool collapse limit (although the factor is four for Ni-Cr alloys). There is not great scope for increased roughing removal rates. The factor increases to the range five to ten for finishing operations. There might be scope for increased removal rates if tool wear could be reduced. The actual removal rates vary some ten-fold from the easiest to machine (Cu and Al alloys) to the most difficult (Ni-Cr and Ti alloys).

For the easiest to machine, tool acceleration and stiffness currently limit removal rates. Limits from this cause are considered in more detail in Section 3.1 and 3.2. In addition to these there are reliability limits. The more severe are the consequences of outlier events (usually tool failures) the more conservative are the machining conditions that are chosen. Reduced variance of tool life could lead to much enhanced productivity where machining processes are not in isolation but an element of a complex process chain (see Section 4.)

2.1.2 Grinding

The mechanics of grinding processes are closely related to those of machining, except that the single cutting tool of machining is replaced by the multiple cutting edges of the abrasive grains bonded into the form of a wheel by a suitable matrix material. Machine design for grinding has similar requirements to that for machining processes, but the distinct limits to process speed in grinding arise from the management of heat and the performance of the grinding wheel over time.
Grinding has a higher energy input than machining per unit of material removed, and most of this energy is converted to heat which is ‘partitioned’ between the workpiece, the wheel and the coolant. Malkin and Guo (2007) report that the high workpiece temperatures that arise when the heat is not partitioned effectively can lead to thermal damage such as workpiece burn (related to accelerated oxide formation), metallurgical phase transformations, tempering and re-hardening, and residual tensile stresses (as mentioned above for machining). The energy partition to the workpiece depends on the wheel and workpiece materials, and the operating conditions, including the delivery of coolant. Howes et al. (1987) note that a water based fluid would boil at 100°C, so will be effective only if the temperature is kept below this point, but for shallow-cut grinding this is unlikely, so the fluid boils and in effect the grinding conditions become dry. Thus, in this case, although the fluids act as lubricants and reduce the energy input, they provide little cooling effect and the workpiece temperature rises rapidly. However in grinding processes with deeper cuts (such as creep-feed grinding), grinding fluids are delivered at high pressure to remove heat, and also provide bulk cooling to reduce thermal deformations. Morgan et al. (2008) show how the achievable flow rate for a coolant supplied from the entry side of wheel (the fraction of the supplied liquid that actual enters the grinding zone) is about a quarter of the jet flow rate, and Baines-Jones (2008) shows how the required jet flow rate is twice the product of the wheel width, the wheel speed, and the mean pore depth in the wheel (roughly equal to mean grain size).

Where tools in machining experience wear, abrasive grinding wheels become clogged by debris, grains lose their sharp edges, and over time as the grains are worn down, more of the matrix binding material is in contact with the workpiece. These effects reduce the material removal rate, and must be countered by the cleaning and dressing processes reviewed by Wegener et al. (2011). For a particular wheel, grinding at double the speed would require double the frequency of cleaning and dressing. However, the efficiency of grinding also depends on the selection of the right abrasive wheel for a given workpiece (as discussed by Savington, 2001) among which vitrified cubic boron nitride abrasives have been particularly important (Neailley, 1988). These have a low
energy partition value which can be attributed to a low specific grinding energy and the high thermal conductivity of the cubic boron nitride grain which enhances heat removal and cooling by the grinding fluid.

Figure 6, which summarises the limits to grinding speed discussed above, shows the importance of heat removal which in turn depends on the rate at which coolant can be supplied to the grinding zone and the design of the grinding wheel. Wheel design also determines the rate at which the wheel must be cleaned and dressed.

![Figure 6. Speed limits in grinding arising mainly from inadequate heat control lead to inaccurate workpiece geometry with various surface defects.](image)

**2.1.3 Electro-chemical machining**

Electrochemical Machining (ECM), a non-contact process, was conceived in the early 1950s to machine advanced aerospace materials such as nickel, chrome, molybdenum and tungsten alloys, which were intractable by traditional machining. Rajurkar et al. (1999) provide an introduction. The metal removal mechanism of ECM is electrolytic dissolution from an anode workpiece, separated from a shaped cathode tool by an inter-electrode gap of 0.1 – 1.0 mm. DC power at around 5-20 V is applied across the electrodes leading to current densities of 50-400 A/cm². An aqueous electrolyte is pumped at a flow rate of 10–50 m/s between the tool and the work, this pumping being required to maintain electrolyte concentration and to remove sludge. Usually the tool is fed towards the work at constant speed to maintain an equilibrium gap while the metal is dissolved to form the desired shape or cavity. Hocheng et al. (2003) report that typical metal removal rates are currently up to around 1.5 cm³/min,
and De Silva et al. (2000) report that ECM can achieve accuracy better than 5 μm and surface finish of 0.03 μm Ra.

Hindjua and Kunieda (2013), in a survey of modelling approaches, state that the rate of ECM processing depends on the rate of electro-chemical reactions, the flow of electrolyte, thermal effects leading to electrolyte heating (and at worst, boiling), and the rate of anodic dissolution. Perfect control of these four features would allow significant speed increases, but in practice production rates are constrained by process instabilities including temperature variation in the electrolyte and the release of hydrogen bubbles at the cathode which lead to variation in electrical conductivity, and in due course, uneven material removal rates. Speed is also constrained by the need to avoid unwanted electrical discharges and short-circuits.

The ECM process is illustrated in fig. 7, but unlike the processes discussed in the previous two sections, the limits to doubling the speed of this process are not shown in the figure, as these limits depend on the maintenance of process stability rather than overcoming any specific barriers. One key feature of this stability is the maintenance of laminar flow in the electrolyte at the highest possible feed rates, to reduce hydrogen bubbling and variations in concentration to a minimum.

![Figure 7. Possible defects arising from electro-chemical machining and key process parameters](image-url)
2.1.4 Laser cutting, drilling and ablation

Under laser irradiation, materials evaporate at low flux or are converted to a plasma at higher fluxes, allowing cutting, drilling or partial removal. The volume of material removed depends on laser power and wavelength and the optical properties of the material. Higher powered laser beams can be used in pulsed mode, to reduce the transfer of heat to the surrounding material (which generally leads to a ‘heat affected zone’ in which material properties are affected). A key attraction of using lasers for material removal is the speed at which the beam can be moved and the relative simplicity of automating laser processes. However their drawback includes relatively low energy efficiency and potentially high capital cost. The limits to manufacturing speed in laser ablation processes can conveniently be grouped into those related to the laser itself, and those related to the workpiece material.

*Limits related to the laser:* material removal through ablation demands high power densities with short interaction times to limit the heat affected zone, and pulsed lasers generate greater power densities than any other heat sources, in the range of $10^{21}$ W/cm$^2$. The power density at the focal point is directly proportional to the square of the original beam diameter and inversely proportional to the square of the beam wavelength and lens focal length, which limits further power increases. The profile of the laser beam (distribution of power-density across the beam) should preferably be Gaussian, so the beam can be focused to its theoretical minimum radius giving maximum power density, but this may not always be possible. The ablated material forms a plasma, some of which will interfere with the future passage of the beam: if the frequency of the light is greater than the plasma oscillation frequency the beam will propagate through the plasma; otherwise, the beam may be reflected, resulting in energy losses. Additional energy losses may occur due to beam absorption in optical components, heat transfer losses due to convection in the shielding gas, radiation from the workpiece, and by heat conduction away from the ablation zone into the workpiece.

*Limits related to the workpiece:* A fundamental limit to laser ablation arises from the reflectance, $R$ (or equivalently, the absorptivity, $A = 1 - R$) of the workpiece,
which depends on both the beam wavelength and the optical properties of the workpiece (Kannatey-Asibu, 2009). The power $q_a$ absorbed by a material, defined as a function of the incident power $q_i$ is given by:

$$q_a = q_i (1 - R) = Aq_i \quad (3)$$

Decreasing wavelengths result in higher absorptivity for metals, whereas absorptivity increases with increasing wavelength for insulators (Steen et al., 2003). Analogously, metal absorptivity increases with increasing temperature and becomes particularly high in the molten state (Ready et al., 2001). The wavelength of a laser beam is defined by the physical properties of the lasing medium employed (e.g. solid-state, gas, dye, semiconductor lasers and so on), and although lasers have been developed for many wavelengths in the ultraviolet, visible and infrared spectra, there are still numerous gaps in the spectrum.

The thermal properties of the workpiece material such as conductivity and latent heat of vaporisation also pose fundamental limits to the laser material removal rate (Ion, 2005). The amount of material that can be removed by vaporisation is limited by the latent heat of vaporisation (Ready, 1997), given by:

$$D = \frac{E_0}{A\rho \left[ c \left( T_b - T_0 \right) + L \right]} \quad (4)$$

Where $D$ is the maximum depth of a hole that may be drilled by a laser pulse with energy $E_0$, $c$ is the heat capacity per unit volume, $T_b$ the vaporisation temperature, $T_0$ refers to the ambient temperature, $L$ is the latent heat of vaporisation per unit mass, $\rho$ the density and $A$ is the spot area of the beam. Materials with higher thermal conductivity (and hence higher diffusivity) allow more heat to be conducted away from the ablation zone, resulting in lower material removal rates (Ion, 2005).

Some heat will always be transferred from the ablation zone to the surrounding workpiece, and a focus for research in laser ablation of metals has been the need to minimise the resulting metallurgical effects in the heat affected zone (Disimile et al., 1998). If the heat transferred is sufficient, this may include melting and re-
solidification to form a recast layer with cracks, changed microstructure and a new oxide layer (Sezer et al., 2006), burrs and spatter (for example Low et al., 2000a), or micro-cracks during laser drilling of nickel alloys (Low et al., 2000b). In severe cases, cracks initiating in the recast layer may extend to the heat-affected zone and into the parent material consequently limiting the fatigue life of the component (see fig. 8.)

![Figure 8. Backscatter micrograph of a laser drilling induce recast, oxide, inclusions, heat affected zone, and micro cracking.](image)

Certain applications, for example the production of cooling holes in turbine blades, require angled holes, and the drilling angle plays a significant role on the development of recast layers, as does the trepanning speed (see fig. 9). Recent trends for avoiding the development of recast layers altogether point into the direction of ultra-fast lasers with a pulse length of femtoseconds (Feng et al., 2005).

![Figure 9. Effects of trepanning speed on the recast layer thickness at a laser beam angle of 30 degrees, peak power of 12.5 kW, pulse frequency of 30 Hz, gas pressure of 100 psi, and hole diameter of 0.8 mm](image)
Laser machining of CFRP (Carbon Fibre Reinforced Plastic) is constrained by significant differences between the constituents’ properties (for example, thermal conductivities of 0.002 and 0.5 W/cm/°C and vaporisation temperatures of 500 and 3300°C for resin and graphite fibre respectively) (Cenna and Mathew, 1997). As a result, heat is transferred out of the drilling zone, melting and vaporising the matrix which leads to defects such as surface damage, matrix recession, craters and delamination. As an illustration, Shyha (2013) observed that a marginally lower beam power of 2250 W, led to a threefold increase in the roughness of CFRP workpieces compared with identical trials performed with a beam power of 2500 W.

Figure 10 summarises the constraints on the speed of laser ablation revealed in this section. Speed limits primarily arise from the power density of the laser, constraints on available laser light wavelengths and the thermal and optical properties of the workpiece.

![Diagram of laser ablation](image)

**Figure 10.** Schematic of laser ablation: how workpieces of various materials respond to pulses of light of varied power, frequency, shape and duration.

### 2.2 Material processing that does not change the mass of components

Casting, forming (or moulding) and heat treatment are all processes that leave the mass of a component unchanged. The two examples here belong to the area of metal forming, which is characterised by large tool forces, a strong dependence on temperature, and the fact that unlike in machining operations,
the workpiece may be affected some distance away from the region of contact with the process tooling.

2.2.1 Hot strip rolling of steel

In the long process chain required to produce steel strip, hot strip rolling generally follows continuous casting and reverse rolling (or roughing). Hot rolling is typically used to reduce the thickness of two-metre wide plates from around 25 to 2 mm as the steel passes through up to five sequential rolling mill stands. The quality of the resulting product depends on it having a specified average thickness (gauge), profile (usually a slight reduction in thickness from centre-line to edge), and grain size, the latter being a function of the strip temperature, the total reduction and the rate at which it was deformed in each stand. In addition to the requirement to meet these quality parameters, the speed of the process is constrained by the need to avoid fast vibrations of the mill stand (chatter), to limit edge cracking and to co-ordinate with the rest of the process chain.

Prior to hot rolling, strips are preheated in a furnace, and during rolling, the plastic work of deformation generates additional heat. Heat is lost outside the mill by radiation and convection to air, and within the mill by conduction to the work rolls (Lassraoui and Jonas, 1991, provide an early analytic model of these processes). Water spray cooling may be used during rolling to cool the work rolls and after rolling to quench the strip. At double the speed, the heat generation in the roll bite will remain the same (the work done is unchanged as the deformation does not change), but the time available for heat transfer out of the strip by all mechanisms is halved, so the strip will become warmer. However, with appropriate spray cooling, the strip temperature can be controlled.

Tekkaya et al. (2015) review current knowledge about property prediction in metal forming, and describe how steel microstructure for a given chemical composition is influenced by rolling: during deformation, the material experiences work hardening, dynamic recovery and dynamic recrystallisation; between the rolling passes, static recovery, static recrystallisation and grain growth occur. Depending on composition and temperature history, precipitation
and phase changes may also occur. As a result Mitao and Yanagimoto (2006) show how by following different thermo-mechanical trajectories, quite different microstructures can be achieved from the same starting material. The dependency of property evolution on histories of temperature, strain and strain rate are strongly dependent on material composition – for example, Yanagimoto et al. (2014) give equations describing the ‘material genome’ of three different steels, in which the rate of dynamic recrystallisation may either increase or decrease as the strain rate increases, dependent on composition. Therefore the effect of doubling speed in hot rolling will be to change the rate of micro-structural evolution processes, which may be beneficial or not, and may be countered by different cooling processes, depending on the specific behaviour of each material considered.

Gauge control in strip rolling is a mature technology: the average separation of the work rolls in each stand is controlled along with the torque applied at each stand, to achieve specified levels of ‘inter-stand’ tension in the strip, and this is well understood and stable, so would be unaffected by doubling speed. The unwanted vibrations of mill chatter, torsional vibrations of the work rolls which appear to be initiated by disturbances to the frictional interface between strip and rolls, would according to evidence gathered by Panjković et al. (2012) be reduced at higher rolling speeds.

The cross-directional profile and flatness (residual stress distribution) in the strip are controlled by adjustments to actuators such as bending jacks on the rolling stand (see for example Montmittonet, 2006), and there is no evidence that this would be inhibited at higher rolling speeds. While edge cracks are an inevitable feature of rolling processes, there is as yet no basis for predicting whether faster rolling would increase their depth, and the subsequent required width of trim.

Finally, a key feature of the process chain for strip metal is its co-ordination: the release of material from the pre-heating furnace must be co-ordinated with the availability of the rolling mill and all downstream processes, as the material flows in a fixed sequence. Therefore there would be no benefit in doubling the
speed of the rolling mill without doubling the throughput of all other processes in the same chain – and as the time for pre-heating is fixed by the geometry of the product, doubling the speed of rolling would require significant investment in doubling the number (or capacity) of pre-heating furnaces.

This section, summarised in fig. 11, has shown that doubling the speed of hot strip rolling could be achieved – the technical constraints mentioned above can be overcome, and the effect of higher speeds on microstructure evolution is so material dependent that no general comment can be made. However, the need to co-ordinate the mill with other equipment in the same chain raises one of the key issues of fig.1: doubling mill speed would only lead to an increase in production speed if all other processes in the same chain operated at double speed also.

![Diagram showing constraints in rolling speed](image)

**Figure 11.** The limits to speed in rolling arise from mill vibration (chatter), heat generation during deformation, and heat transfer between strip and rolls.

### 2.2.2 Incremental sheet forming processes at room temperature

The high cost of tooling for high-speed mass production of sheet metal components such as car body parts has driven a wave of interest, led from Japan (Allwood and Utsunomiya, 2006) aiming to develop flexible production processes which can form many different components without specific tooling. The earliest metal forming processes, exemplified in the village blacksmith, had this flexibility, and other craft processes such as metal spinning, the English
Wheel and the Power Hammer are still in use for part customisation and in low volume sectors such as aerospace. The process of single point incremental forming (Jeswiet et al., 2005) has attracted the greatest attention from researchers, but as yet had little industrial take-up, due to its poor accuracy and high residual stresses.

These processes mainly operate at room temperature, and due to the need to move tools repeatedly over the workpiece surface, they operate at relatively low speeds. As a result the forming limits of the workpiece – whether arising from local or global necking, or from wrinkling – are unaffected by process speed, which is instead determined by the rate at which the tools can accelerate relative to the workpiece. Metal forming generally leads to higher tool forces than machining operations, so the equipment must be stiff and the workpiece held firmly. The need to have the tools pass repeatedly over the area of the workpiece thus requires cyclic acceleration and deceleration of several axes of movement which themselves have significant inertia, but experience (Allwood et al., 2005, Music and Allwood, 2011) has shown that because of the use of high ratio gearboxes between motor and linear axes, the rate of acceleration is typically constrained by the moment of inertia of the motor rotor.

Figure 12 summarises this simple deduction, illustrated with a single point incremental sheet forming process, where tool acceleration and vibration associated with rapid changes in acceleration (jerk) are the main constraints on process design: the limits to increasing the speed of incremental sheet forming processes are related to machine design, and not process mechanics.
2.3 Material processing that adds to the mass of a component

The two processes in this section build up a component additively – from droplets of liquid in the first case, and sintered particles of powder in the second. Both are relatively new techniques being developed for specialised high value applications.

2.3.1 Inkjet printing

Inkjet printing is not commonly thought of as a manufacturing process, but it is of increasing interest as a method of depositing small droplets of liquid materials under precise positional control. It is now being used not only for printing text and graphics but also as an enabling technology for the manufacture of thin-film electronic devices, displays and sensors and in certain processes of 3D printing (Castrejón-Pita et al. 2013). In inkjet printing the ink or other liquid is delivered on to a substrate in the form of small droplets, typically 10 to 100 µm in diameter. For high-precision applications the drops are usually generated by a printhead which contains an array of individual nozzles (typically 100 to 1000 in number), each of which can be addressed separately by an electrical signal. Commonly these ‘drop-on-demand’ prinheads for industrial application use piezoelectric ceramic actuators to eject the liquid through the nozzles; the elongated jets generated in response to the electrical impulses collapse in flight under surface tension forces to form individual drops (Hutchings and Martin, 2013). The design of the printhead and the size of the nozzle determine the diameter $d$ of the drops and the maximum frequency $f$ at which they can be generated. Figure 13 shows the performance of typical current commercial piezoelectric inkjet printheads, from five different manufacturers and with a range of different design architectures, with drop volume plotted against maximum printing frequency. There is considerable variation in performance between the printheads from different sources, probably associated with differences in the proprietary internal designs of the ink channels and actuators, although further variation in performance can also result from the rheological behaviour of the fluid from which the drops are formed.
Although there is a wide range of drop volumes (from 5 to 200 pL, corresponding to spherical drop diameters from 10 to 36 µm) and frequencies (from 4 to 50 kHz), there is a general trend for the larger drops to be produced only at lower frequencies; the maximum volume delivery rate per nozzle varies rather less, from about 0.2 to 1.5 µL/s for all these cases.

To print a pattern, the printhead and substrate are moved relative to each other while the desired sequence of printing instructions is fed to each nozzle. The speed at which inkjet printing can produce a specified pattern therefore depends on the maximum size of droplet that can be used, the frequency of droplet ejection and the speed of printhead movement over the substrate.

The droplet size plays a key role in determining the resolution with which printed features can be defined. For precise printing with high resolution, small drops are desirable, but as the drop size is reduced aerodynamic effects become increasingly important. Current drop-on-demand printheads eject drops at speeds between 5 and 10 m/s and typically operate at a standoff distance from the substrate of 0.5 to 1 mm, which is needed so that the ejected jets collapse satisfactorily into drops. Air drag on a 1 pL drop (12 µm diameter, the smallest
currently used even in desktop printers) initially travelling at 5 m/s will halve its speed over a distance of 1 mm. This suggests that the practical lower limit for drop size has already been reached. When the liquid drop strikes the substrate it will spread out to form a circular deposit: a ‘dot’ of ink in the case of graphical printing. The diameter of this dot depends on the behaviour of the liquid in contact with the substrate; for a non-porous substrate it will be governed by the final contact angle, and will typically be about twice the diameter of the droplet which formed it (Jung and Hutchings 2012).

The maximum frequency with which drops can be ejected from a drop-on-demand printhead is determined by the mechanical design of the printhead (which for example will lead to certain resonant frequencies in both the mechanical structure and the internal liquid-filled region), and also by the behaviour of the ink once it has left the nozzle. It is possible that with new designs of printheads and by formulation of the ink to optimise its rheological behaviour (which will control the way in which the ligament of ink which follows the main drop merges into it), further increases in practical jetting frequencies may be achieved.

The maximum relative speed of movement which can be achieved will depend on the nature of the pattern being printed: to achieve a sequence of dots from a single nozzle so that they just touch each other in the direction of motion (for example, to print a continuous line of colour, or an electrical conductor), that speed cannot be greater than ~2df. For all the printheads in fig. 13 this maximum speed lies between ~0.3 and 1.3 m/s.

Drops can be formed at higher frequencies by printing in the continuous inkjet (CIJ) mode (Martin and Hutchings 2013) which is commonly used in industry for simple coding and marking purposes. The mode exploits the natural tendency for a continuous liquid stream to break up under surface tension forces into separate drops; the wavelength at which any disturbance will grow most rapidly is about 4.5 times the jet diameter, and most industrial CIJ printers stimulate breakup at this wavelength by applying a periodic disturbance to the jet at an appropriate frequency (the Rayleigh frequency). Under these conditions the
drops have a diameter of 1.89 times the initial jet diameter. To use the CIJ method as a practical printing process, individual drops must be steered to the correct locations on the substrate, or deflected so that they do not land on the substrate at all; in most current systems this is achieved by inducing an electrical charge on each drop, which then passes through a transverse electric field. The continuous lines in fig. 13 show the performance of a CIJ printer operating at the Rayleigh frequency for two different jet velocities, chosen to cover the range used in drop-on-demand printheads. It is clear that CIJ technology can in principle achieve much higher material deposition rates than current drop-on-demand printing, but at present it lacks the versatility and flexibility to be used more widely as a manufacturing process. A major obstacle is the challenge of producing an array of many CIJ nozzles, although some commercial designs are available (Martin and Hutchings 2013).

Figure 14 summarises the constraints on the speed of inkjet printing: the rate of material flow depends on drop size and ejection frequency, with smaller drops preferred for precision, but limited by aerodynamic effects, and higher frequencies preferred to allow faster printhead motion, but constrained in theory by the Rayleigh frequency, and in current practice by resonances and the time needed to refill the nozzle chamber in the drop-on-demand printhead.
The limits to inkjet printing speed relate to the behaviour of each drop: ejection frequency, behaviour of the drop in flight, and effect of impact as the drop hits the target substrate.

Figure 14. The limits to inkjet printing speed relate to the behaviour of each drop: ejection frequency, behaviour of the drop in flight, and effect of impact as the drop hits the target substrate.

The question of whether inkjet printing has reached its limit of speed has several dimensions (Castrejon-Pita et al. 2013). Current maximum substrate speeds of about 1 m/s for inkjet printing are substantially lower than those used in more conventional printing processes (5-10 m/s). Higher ejection frequencies demand more rapid refilling of the nozzle chamber, as well as careful management of oscillations within the printhead structure and the fluid column. This has implications not only for printhead design and operation, but also for rheological design of the printing fluid. Higher print speeds demand greater rates of data transfer to the printheads. Ultimately, the limit to speed may well be controlled by the level of error acceptable in the final product, with material deposition for a structural component in some cases being more tolerant of defects than a more demanding electronics application.

2.3.2 Additive manufacture by laser sintering of polymer powder

In laser sintering processes, a part is built up incrementally in a bed of thermoplastic powder: a laser beam tracks a chosen path over the top of the powder bed to cause partial melting and fusion of the upper layer of powder; the build area drops by one layer thickness (often ~100μm), a new upper layer of powder is added and the process repeats. The quality of the resulting part is defined by its density, the degree of particle melt, its surface roughness, and its crystalline content which is largely dependent on cooling rate; for example, Zarringhalam (2007) has shown for Nylon-12 that an increase in cooling rate from 1°C to 23.5°C per minute causes an increase in Elongation at Break from 13.5% to 19.1%, and an associated decrease in Tensile Strength from 42.5MPa to 40.6MPa.

Production of a part by laser sintering requires pre-processing, sintering and powder addition and post-processing, each of which affect overall processing speed.

In pre-processing, the powder may be sieved (for uniformity), have static electricity removed, be dried and pre-heated. Pre-heating is required to minimise the thermal gradient between sintered and un-sintered regions, to
minimise the laser energy required to achieve melting, and to reduce ‘curling’ of parts during a build. Figure 15, for example, shows how a semi-crystalline polymer such as Nylon-12 (currently the most commonly used polymer for laser sintering), must be warmed to just below its melt temperature, but above its re-crystallisation temperature, prior to laser scanning. The larger the difference between its melt and crystallisation points (known as the super-cooling window), the easier the material is to process, and Tontowi and Childs (2001) have shown how temperature variations throughout the powder bed affect part density, and may prevent building in cooler (often outer) regions of the build chamber. This in turn may prevent optimal utilisation of the full build volume, leading to higher cycle times per part. The requirement for precisely controlled uniform temperature throughout the powder bed prior to laser sintering requires time which is fundamentally limited by the size of the bed and the thermal conductivity of the powder.

![Figure 15. ‘Ideal’ super-cooling window for ‘good’ semi-crystalline laser sintering polymer](image)

During laser sintering, effective transfer of energy into the base powder is crucial. Starr et al. (2011) show the effects of energy-melt ratio, defined as the ratio between the energy input into a layer of powder, and the theoretical energy required to fully melt the layer. Energy input is affected by several factors, including beam speed, spacing between scans and layer thickness. The energy required to melt a layer is dependent on melt temperature, packing density, heat of fusion and specific heat, so therefore material design to allow modification of
any of these parameters could enable processing at lower energies and thus faster speeds. The Frenkel model of viscous sintering reported by Muller et al. (2008) can be used to predict sintering rates for polymer powder.

\[
\frac{x}{r} = \sqrt[2]{\frac{3\gamma t}{2\eta}}
\]

(5)

where \(x\) is the neck radius between two particles, \(r\) the original radius of one particle, \(\gamma\) the surface tension, \(\eta\) is viscosity, and \(t\) the sintering time. At lower viscosities, sintering time will be reduced (accompanied by a negative effect on geometric accuracy if the molten powder flows too freely), suggesting a benefit to the higher temperatures caused by increased energy input. However, as with other polymer processing methods, degradation will prevent large temperature increases. The term Stable Sintering Region is used by Vasquez et al. (2011) to describe the window between a material’s melt and degradation temperatures, within which high mechanical properties can be achieved. Increasing scan speed and spacing between scans can reduce the time taken to produce a part: however, the resultant decrease in energy input can lead to a decrease in Degree of Particle Melt (DPM), a measure of how much of the powder is actually melted during the scanning process of laser sintering. Majewski et al. (2008) have shown that this in turn leads to a decrease in mechanical properties. The size of the powder bed and part orientation within it are also important: Caulfield et al. (2007) demonstrated the effect of build orientation on part properties; for example parts built vertically consistently exhibit density reductions of more than 10% when compared with those built horizontally. Reductions in this anisotropy would allow optimum build arrangements for increased production speed. Bacchewar et al. (2007) have also demonstrated that layer thickness and part orientation additionally affect surface roughness, with parts built vertically or at higher layer thickness exhibiting higher surface roughness.

Following the build, the parts must be cooled while encased in un-sintered powder. Both the heating and cooling stages add considerable time to the process (warm-up times of 2-3 hours, and cool-down times similar to the length of the actual build process are typical). The thermal conductivity of polymers is
known to be low (generally 0.1-0.6W/mK, (Yang, 2007)), which will necessarily increase the times associated with these stages. When considering a powder system, Shapiro et al. (2004) have shown that the size and shape of the particles will have a crucial effect. Thermal conductivity within the gas-filled gaps between particles is poor, meaning a particle distribution conducive to good packing is essential.

Figure 16 summarises these limits to doubling the speed of laser sintering of polymer powders: the pre and post-processing steps are primarily constrained by the need for uniform temperature changes in a material with low thermal conductivity, and the production step is constrained by the need to add sufficient heat to cause a high degree of particle melt while avoiding any part of the particle reaching the polymer degradation temperature.

Figure 16. The speed limits to laser sintering of polymer powders include the process itself, and the operations that must occur before it begins and after it completes.

While each of the three stages in the Laser Sintering process illustrated in fig. 16 can limit the speed of part production, many of the pre- and post-process stages can be automated or conducted offline. Although, manufacturers have made only a few attempts at such automation to date, future development is likely. The part-build stage is therefore key to future speed increase.
The single-line scan process illustrated in fig. 16 may be close to its speed limit: incremental increases in laser speed and absorption efficiency may be possible, but are unlikely to create step change improvements. New process designs aiming to increase the instantaneous area of sintering are being developed and will be discussed in section 5. The remaining constraint on the speed of sintering is therefore the characteristics of the raw materials. Properties such as viscosity, thermal conductivity, and thermal degradation characteristics constraint current speeds and recent interest in this area from global material manufacturers is likely to lead to near term improvements.

2.4 The operation of manufacturing systems

The physical processes discussed in sections 2.1-2.3 do not operate in isolation, but are part of a larger system. The discussion has already referred to links between physical process speed, and that of surrounding processes – such as the pre-and post-processes of laser sintering, or the pre-heating furnaces of hot rolling. This section considers three aspects of the wider system – two forms of metrology, and some features of the ergonomics of assembly operations.

2.4.1 Metrology for surface topography

The performance of high-value devices such as biosensors, holographic security features, modern projectors, touch-screen devices, functional packaging and high-density circuitry made in high-volume device can depend on management of a range of defects, but assessment of their quality is increasingly outside the capabilities of available metrology. Figure 17 shows two examples: the light collection optic of fig. 17a incorporates dense surface structures with high aspect ratio geometries whose precision strongly influences collection efficiency. Figure 17b shows a master stamp for micro-optics being produced by a drum diamond turning operation. The process takes several days but the cost of checking the surface with current metrology is too high to allow individual product verification.
Figure 17. Examples of devices whose high value can only be confirmed with high precision surface metrology: (a) a photovoltaic panel (b) a master stamp for micro-optics. (images courtesy of National Physical Laboratory, UK)

The challenge for measuring surface topography is illustrated in fig. 18: existing metrology solutions trade-off sample-area throughput against resolution, but manufacturers of products such as those in fig. 17 would like to measure features accurately over large areas to very high spatial resolution.

Figure 18. The challenge of surface metrology development

The most accurate and precise forms of surface measurement come from tactile probes, which must be scanned across the surface at slow enough speeds to avoid distortion due to the dynamics of the probe. Non-contacting, largely optical, systems therefore show more potential for sensing over larger areas. There are two distinct classes of instrument: (a) those that measure over large areas
(square metres) with tens to hundreds of micrometres spatial resolution (for example, fringe projection, photogrammetry and Moiré interferometry) (Harding 2013), and (b) those that measure over small areas (up to a few square millimetres) with spatial resolutions of the order of a micrometre (for example, coherence scanning interferometry (CSI), confocal microscopy and focus variation microscopy (FVM)) (Leach 2011). Essentially, the former class is camera-limited, and the latter is objective-limited. There have been several attempts to try and combine the two classes (see for example, Kayser 2004), but as yet this approach is not ready for use in high-speed manufacturing.

A simple calculation, considering only the measurement acquisition rate, underlines the problem for imaging microscopy of features on a flat substrate. Inspection throughput equivalent to 1 m\,min\(^{-1}\) of a one metre width web, implies a minimum inspection rate of 0.016 m\(^2\)/s. Using a 5\(\times\) objective (numerical aperture (NA) of 0.15, Rayleigh resolution limit 5 \(\mu\)m), measurements must be completed at a rate of 4000 fields of view per second. For a 100\(\times\) objective (NA of 0.95, Rayleigh limit 0.25 \(\mu\)m), the minimum rate is on the order of 2 \(\times\) 10\(^6\) fields of view per second. The additional axial scan or equivalent operation associated with all topography techniques will increase these frame rates (and associated data flow) by perhaps three orders of magnitude, therefore, making this type of measurement infeasible.

The challenge of doubling the speed of surface metrology is therefore to find a means to do so without compromising on the accuracy of the resulting measurements.

2.4.2 Thermal imaging

Thermal imaging can be used for on-line process monitoring, and non-destructive testing of finished parts or of parts in service to estimate their remaining useful life. An introduction to the topic by Maldague (2002) and a review of applications by Meola and Carlomagno (2004) describe the basic principles of thermal imaging: infra-red radiation is emitted by all bodies above absolute zero, with radiative power peaking at wavelengths that decrease with the body’s temperature; a camera which can identify this peak wavelength can
therefore be used to predict its temperature. Maldague (2002) distinguishes passive thermography (where the body being examined is already above ambient temperature, for example during a hot manufacturing operation) and active thermography (where an external stimulus is required to induce thermal contrasts.) An example of this second approach is shown in fig. 19, being used to inspect a part in service for sub-surface cracks: a burst of heat energy is applied to the component, and anomalies in the resulting history of temperature across the component surface may reveal sub-surface features. For instance, a small void would have a different heat transfer coefficient from the surrounding material, so might heat more slowly – creating a ‘cool spot’ on the surface.

Figure 19. Automated Thermographic inspection of a turbine blade supported by a robot (Mehnen et al., 2014)

Thermal imaging cameras can operate at high frame rates and with fine resolution. However, because the surface temperature is a function of the history of heat transfer into, within and out of the inspected component, thermal image processing is challenging. Images from un-cooled thermal cameras can be noisy, so a minimum feature size of 2 to 3 pixels is often recommended to classify a feature as a valid detection (Maldague, 2001); in later frames in the time sequence of images, features that are deeper into the product will appear smaller than they are (Almond and Lau, 1994); because of heat dissipation through the material, images will appear blurred.
Limits to the speed of thermal imaging are determined by the thermographic properties of the inspected material and the equipment. To avoid heat building up in the equipment after each "pulse" of heat is delivered, the heat source may require time for cooling. In the system illustrated in fig. 19, a single shot lasts only for milliseconds, but the safe cooling interval is about 5 seconds.

2.4.3 Ergonomics in assembly operations
Since the days of Henry Ford, increases in speed in assembly operations have been associated with the division and specialisation of labour, and for assembly line workers this has led to the need to repeat highly proscribed actions at a rate defined by the progress of the assembly line. Doubling the speed of assembly would require further specialisation, and the limits to human performance in this context depend significantly on the way that the design of the task is physically comfortable or not. Even with on-going increases in the number of tasks that can be automated, it is unlikely that human work will ever be eliminated from manufacturing, so human limits to performance may always be a critical bottleneck for further speed increases.

Ergonomic design methodologies are usually based on a single performance measure (i.e. lift index, work postures, etc.) related to standards such as the Ovako Working Posture analysis System (OWAS) or the equations of the National Institute for Occupational Safety and Health (NIOSH). Lin and Chan (2007) and Waters et al. (2007) report examples of such design methodologies based on a single ergonomic performance measure. The integration of two or more performance measures into design methods such as in Andreoni et al. (2009) and Baines et al. (2005) is claimed to lead to better task design. Battini et al. (2011) propose a methodological framework that takes into account technological variables (related to work times and methods), environmental variables (absenteeism, staff turnover, work force motivation) and ergonomics standards. Several authors provide data on the relation between ergonomics and work performance. For example, Thun et al. (2011) analysed the use of ergonomic principles in task design in 55 companies in the German automotive industry, and report improvements in productivity, flexibility and reduced absenteeism as a result of good practice.
A significant area of research has aimed to develop modelling tools for use in conjunction with assembly line and product design tools, to include task design features such as motion efficiency, operation safety and productivity as part of a holistic view of new product introduction. Wilson (1997) provides an early overview of attributes and capabilities of virtual 3D environments devoted to support ergonomic design including tools such as Digital Human Models (DHMs) and Virtual Reality (VR) techniques to evaluate the performance of the manufacturing systems can influence. As an example of the state of the art, Lamkull et al. (2009) report on the evaluation of 155 manual assembly tasks with a car assembler in Sweden, comparing the predictions of a DHM based tool with the evaluation of six in-company experts. The computer based tool correctly identified several postural and movement problems, but failed to identify further problems associated with balance, field of vision and forces required to complete tasks.

Key technologies such as motion capture, haptic modelling and rendering are currently under development to add more realism to virtual environments, overcoming the lack of feedback on forces and allowing users to carry out reachability and visibility analysis. The challenge of integrating ergonomics into CAD tools used to plan assembly depends on characterising and encoding biophysical performance limits of the full range of human activities likely to be involved in assembly. Future research and development in the development of advance virtual reality technologies is likely to aim at creating more intelligent sensors that can capture information beyond motion, such as human emotions (for example stress level) that can be used as part of the feedback to task design.

2.5 The management and co-ordination of manufacturing
Beyond the operation of the manufacturing system discussed in the preceding section, manufacturing further requires overall co-ordination – internally between processes, within the supply chain to manage costs and cash flow, and in a wider sense with customers through the on-going process of design and new product introduction. These three aspects are considered in this section.
2.5.1 Job shop scheduling

A manufacturing system making one unvarying product to satisfy a steady rate of demand can be optimised to achieve perfect utilisation of all resources. However, outside the world of steady state chemical processing (such as petro-chemical refining, paper making, or glass production) this simplicity is denied by customer preference for variety, and production engineers must address the challenge of scheduling work to optimise throughput and capacity utilisation, knowing that perfection is impossible. Koren et al. (1998) give a simple illustration of this, showing the influence of process layout on production of a single product family – with productivity reducing as the number of sequential operations increases. Adding machines to an existing production system could also result in interoperability costs (Newman et al., 2008). For example, the programs written to enable manufacturing on an existing machine would not necessarily produce the desired effects on a new set of machines, and new programming would add to the cost of the speed increase (Nassehi et al., 2008).

In many practical cases, expanding the capacity (speed) of a manufacturing system will also be combined with increasing the variety of products produced. Research has shown that with an increase in variety, productivity decreases and thus doubling the machines would not result in doubling of capacity when product variety increases as well. Responses to the challenge of managing complexity related to product variety have, as reported by ElMaraghy et al. (2013) largely focused on designing products in ‘families’, or ‘platforms’ where the process of manufacture can be least disrupted as optional features or overall designs evolve with time.

Extending this problem of complexity out of a single firm and into the supply chain, requires the addition of new inventory and storage (Dallery and Gershwin, 1992), excess capacity to cope with co-ordination delays (Deif and Elmaraghy, 2009) and changed supply-chain logistics (Gunasekaran et al., 2001).

The cause of all the issues raised in this section is variety. In the absence of variety – if every task for every product was always completed in the same time, without variation – doubling the speed of a manufacturing facility could be
achieved simply by doubling the number of machines in it, without any of the physical process speed increases discussed in section 2.3. However, in the reality of variety – in the performance of any given task, due to errors, from product variety or supplier unreliability – disturbances propagate through the system, and doubling the speed of the system would require more than doubling the productive elements within it.

2.5.2 Cost estimation

The introduction of new products, or the quotation of prices to customers for non-standard production, depends on cost modelling. Cost modelling is typically based on concrete data and design models. If all the information required for calculating the costs is given, the time to generate a cost result is determined solely by the time required for calculation. However, if the data is not all available and only imprecise design information is available, cost estimations must be made based on human reasoning, and this takes time.

Houseman et al. (2008) and Coley et al. (2007) investigate the human cognitive processes of cost estimation, using techniques such as ‘thinking aloud’ and protocol analysis to explore fundamental cognition behaviour and to compare the performances of experienced and novice cost engineers. Part of this analysis, illustrated in fig. 20 shows how complex the decision processes are. Five expert cost estimators from Rolls Royce, the European Space Agency and the Ford Motor Company estimated the cost of producing a mountain bike, using only a pen, paper and a calculator, while saying aloud everything they thought about while carrying out the task. Their statements were recorded, and transcribed using a coding scheme for their cognitive actions, divided among four features: macro-strategy (e.g. top-down or bottom-up), micro-strategy (such as consulting external information, or making an assumption), problem domain (e.g. materials, processes) and level of abstraction (e.g. whole system, subsystem, etc.) The figure which is representative of the trials shows how one of the participants swapped repeatedly between various micro and macro decisions trying to capture and integrate structural design options as well as estimating cost information while swapping between various problem domains.
The time the cost estimators devoted to activities categorized as ‘analysing the problem’ was surprisingly low, at 7%, with almost 20% of the time spent employing established techniques. 10% of was used for personal estimating strategies and up to 14% for writing notes. The results of this study suggest that humans work efficiently when combining experience and estimates when dealing with cost engineering questions, but that the speed of their activity is also limited.

In parallel with the discussion on biophysical performance and ergonomics above, this section raises the question of whether it is possible to double the speed of human activities such as cost estimation. The analysis by Houseman et al. (2008) and Coley et al. (2007) suggest that the fundamental limits of the speed of determining cost lie in the speed of access to and assessment of concrete data as well as in the time required for communication when collaborative cost assessments are required. Increased data availability and accuracy can improve the accuracy of the final results. However, asking the humans to double their speed in assessing data will lead to increased uncertainty in the outcomes. In parallel with biophysical limits, this suggests that doubling
the speed of manufacturing must also consider the cognitive limits of human performance.

2.5.3 Design

Design unlike manufacturing is a non-mechanistic, non-deterministic process without a defined set of specific rules and as such can be difficult to specify and control, particularly when multiple stakeholders are involved. Design has been described by some researchers as being behavioural rather than mechanistic and as such the personal view and opinions of the designer plays a critical role (Cross, 2004), (Ahmed et al., 2003). This impacts directly in the way designers view the process of design and creativity and ultimately the speed at which to execute the design process. Whilst creativity is at the forefront of the designer's rationale, innovation is considered to be more critical as this leads to a tangible revenue stream. Creativity however, is a major factor for successful design and is a key driver for innovation. It can also enable a company to develop new products better and faster (Jones et al., 2014).

There is little evidence to suggest an absolute or theoretical limit to design speed, as this is a multifaceted complex procedure that depends on the product being designed. Instead research has been conducted on providing better mechanisms to enable more efficient design, consisting of such approaches as 'Design for X', Axiomatic design (Suh, 1998), robust design (Taguchi, 1993), using a signal to noise ratio to measure the robustness of a product, and systems design (Ritter et al., 2014).

The process of generating a new product is typically carried out in a number of key stages as described by design processes such as that of Pahl and Beitz (2007). These process fundamentally rely on an iterative strategy requiring a number of design and test iteration phases. Whilst this provides a clear product development strategy, it inhibits increases in design speed. In fact, the main constraints to increasing design speed relate to the monetary value of the product and the synergies involved in the design process between the varying stakeholders, such as the client, the design team, product managers, project managers, assembled specialists etc. The need to co-ordinate among these
stakeholders is one of the major limiting factors in reduced design speed, and with increased monetary value, the number of stakeholders increases. Bailetti et al. (1998) discuss the issues of coordinating different stages within a product design. Their findings suggest that coordination challenges during the design phases will invariably change during the product development life cycle and reach a maximum at the midpoint of development.

A particular industrial example of design occurs in the development of the fixturing which is used to position, orient and hold a workpiece relative to manufacturing process machinery. The workpiece, which is new for each new product, can be arbitrarily variable in its geometry but the lay-out of the processing machinery is generally fixed and/or standardised. As a consequence, it is generally necessary to design and manufacture a bespoke fixture for each workpiece, but this reduces manufacturing speed. Bi and Zhang (2001) review research addressing this problem, giving priority to two key themes: a research community has grown-up investigating the use of Computer Aided Fixture Design (CAFD) (Wang et al., 2010) with approaches ranging in complexity from simple systems where an experienced (human) designer controls the placement of pre-built modular blocks to intelligent automated design systems based on process forces and system dynamics; other researchers are aiming to develop conformable fixtures which can be applied across a wide variety of parts. These include ‘Reference Free Part Encapsulation’ systems which use phase-changing substrates (such as wax or low-melt alloy) to hold the component during processing (Sarma and Wright, 1997), conformable "pin-board" type fixtures (Hurtado et al., 2002), swarm robotics based fixtures (Zielinski et al., 2013) or adhesive clamping (Raffles, 2013).

These approaches to innovation in fixturing – based on reducing the size of the solution space through modularity or by limiting innovation to a small number of ‘conformable’ features – may also suggest a possible route to increasing the speed of more general design problems.
2.6 Key themes in the pursuit of manufacturing at double the speed

The case studies reported in this section are not by any means a complete catalogue of manufacturing activities, but were intended as a representative sample. Exploring the constraints on doubling their speed has revealed several common themes, which can now be identified in order to give structure to the next section which will examine fundamental limits to key aspects of speed in manufacturing.

The discussion of materials processes revealed four common concerns: limits to machine stiffness affect both the static precision of manufacturing, and also are the key to reducing unwanted vibrations, whether chatter in machining or metal rolling, or nozzle vibration in inkjet printing; the speed of processes such as machining, incremental sheet forming, or point metrology in which a localised actuator must travel in a complex path over a workpiece, are constrained by axis acceleration limits; the most common constraint relates to the speed at which a workpiece (and sometimes a tool) can be heated or cooled to a specified uniform temperature, as discussed for pre and post-processing of polymer powder sintering, for active thermography and for the electrolyte in electro-chemical machining among others; the delivery of fluids to the interface between workpiece and tool, for cooling or lubrication, and the flight of ink to the substrate, depend on achieving maximum fluid flow rates. Therefore section 3 will consider the fundamental limits to machine stiffness, axis acceleration, heat transfer and fluid flow, to allow exploration of opportunities to improve speed in these areas.

Some of the discussion in this section revealed speed constraints related to equipment – the need for higher power density in laser beams, or the need for greater precision in optical metrology, for example – but these developments are specific to individual processes, so are excluded from the next section as they are better discussed in existing specialist communities. Two features of materials science have also been discussed. Firstly, continuous developments in material design support manufacturing at higher speed – for example through development of new tools in machining or abrasives for grinding, and this development will inevitably continue: the benefit of creating polymers that allow
a wider stable sintering region, for example, is likely to motivate appropriate research in polymer formulations. Secondly, material scientists are continually working to improve their ability to predict material properties as a result of processing. The discussion on hot rolling of steel in particular demonstrated that as this understanding develops further, new options for speed increases will become apparent – for example if a significant increase in strain rate allows the creation of new or equivalent microstructures at lower temperatures than those used today. Both of these material science developments are excluded from the next section, as they are already under intense development within the materials community.

The discussion of the context of manufacturing processes as part of a wider system extending from process co-ordination out to the need to develop new products that meet evolving customer preferences revealed three common features that will be discussed in section 4. A common theme across much of the discussion was about the efficient co-ordination of tasks with variable, and often unpredictable, durations. Whether in co-ordinating the equipment in a job shop, or the many stakeholders of a design process, the limits to speed increases depend on developing means to use resources efficiently despite uncertainty, and this is a key theme for section 4. The discussion on ergonomics raised the question of the bio-physical limits to human performance, and both in considering design and cost estimation, the question of limits to human cognition was raised – and both these human limits will also be considered.

3. Speed limits for material processing technology

The conclusion to section 2 identified four key fundamental limits to the speed of physical manufacturing processes, each of which is discussed here. A representative simple analytical model of each feature is derived to allow identification of key constraints to speed increase, allowing comparison with current practices.
3.1 Machine stiffness

Several processes discussed in section 2 are constrained by machine stiffness – either for geometrical accuracy or for the avoidance of unwanted vibrations. Figure 21 illustrates a representative stiffness problem in machine design – with a compressive force acting within a machine frame.

![Figure 21. A representative machine stiffness problem: a pin-jointed frame comprising four solid square-sectioned beams, loaded internally by the opposed forces of some manufacturing process.](image)

Assuming that the frame must not intrude into the working area (LxL), the frame is to be designed by varying the square section of the four members of the frame (TxT) so that the opposing points of application of the compressive force $F$ both deflect by some amount $\delta \leq \delta_{\text{crit}}$. The deflection arises from the stretching of the vertical members and bending of the horizontals, so assuming that the approximations of simple beam bending theory apply,

$$
\delta = \frac{F}{2ET^2} \left( \frac{L+T}{2} \right) + \frac{F(L+T)^3}{48EI}, \quad \text{where} \quad I = \frac{T^4}{12}
$$

$$
= \frac{F}{4EL\beta^2} \left[ 1 + \beta + \left( \frac{1 + \beta}{\beta^2} \right)^3 \right], \quad \text{where} \quad \beta = \frac{T}{L}
$$

(As the section of the frame $T$ increases, the distance between the pin joints increases as $L+T$.) Apparently as $T$ increases, both terms in (6) will tend to zero, however, at some point the local surface indentation at the point of application of $F$ will dominate the bending of the horizontal members. This indentation should
be calculated numerically in reality, but an approximate asymptote to its effect can be constructed using the Hertzian theory for a stiff sphere of radius $R$ indenting an elastic half space, which leads to deflection,

$$\delta = \sqrt{\frac{9F^2(1-\nu^2)}{16E^2R}}$$  

(7)

Figure 22 shows the deflection calculated by (6) and the asymptote of (7) for a machine with a steel frame, with a working area of 1 m$^2$, subject to a compressive force $F$=1, 10 or 100 kN applied to the frame via a hemispherical-tipped shaft of diameter 10cm.

![Figure 22. Machine stiffness as a function of frame member design (solid lines), with an asymptote for local indentation due to the applied force (dashed lines) for three representative forces.](image)

For the particular geometry and material specified, fig. 22 shows that the frame stiffness increases as the thickness $T$ increases – the stiffness increases around 1000 times as the bars increase from 2cm square-sections to around 15cm. However, beyond this, there is no merit in further increases in the bar section, as the local indentation caused by the application of the load $F$ to the frame creates more deflection than the frame. For example, when the load is 10kN, the deflection cannot be reduced below around 0.03mm.
The situation in fig. 22 is hypothetical, and more sophisticated modelling of machine static and dynamic performance is possible, as amply demonstrated in the overview by Altintas et al. (2014). However, fig. 22 usefully illustrates the fact that machine stiffness cannot be increased indefinitely by the addition of more material to the frame. The linear dependence of stiffness on Young’s Modulus in eq. (6) suggests that some increase in stiffness can be created by alternative material selection; the most advanced technical ceramics have around double the stiffness of the steel used in this example. However beyond this, a different approach is required – and is in fact used already in rolling mills.

To cope with the very high forces involved in rolling, mills are designed with “screws” that adjust the spacing of the outer (‘backup’) rolls, to allow control of the average thickness (gauge) of the strip, even as the force required to deform it varies, for example as the incoming strip temperature varies. In this case, the adjustment can operate relatively slowly, so the demands on the actuator are relatively easy to meet. More generally, the option to substitute actuation for stiffness depends on a comparison between the dynamics of the available actuators and the rate of change of the applied load, F. Huber et al. (1997) explore the availability of actuators, parameterised by their maximum actuation strains and stresses, stiffness and temporal frequency, to provide a basis for the exploration of this area – and both the use of actuation in this sense, and the development of new actuators for this purpose is an open opportunity in manufacturing.

Alternatively, for the high frequency unwanted vibrations associated with chatter for example, the machine design can be adjusted either with material selection to favour damping within the frame, or for further opportunities to develop active or passive damping via additional components. An interesting and as yet unexplored opportunity in this area is created by the work of Smith and Wang (2004) on the “inerter” as a novel component in passive dynamic systems.

3.2 Axis acceleration

Several of the processes in section 2 which depend on scanning a point actuator or sensor across a workpiece were fundamentally constrained by the speed of
scanning which, for all except the most simple cases, is constrained by the acceleration of the point tool. The ‘jerk’ of the tool is a function of how well the rate of change of acceleration can be controlled, so is a feature of the control system more than a limit. Reports from designers of CNC tools suggest that this limit is frequently related to the inertia of the rotor in the motor driving each axis rather than from the tool itself, and fig. 23 shows, for data from the catalogue of a contemporary motor manufacturer how the maximum acceleration of a motor decreases with its maximum torque – the larger rotor required for higher torques has a larger moment of inertia, so limits acceleration.

![Figure 23. Data from a current manufacturer on motor performance](image)

Figure 23 proposes a simplified model of a motor driving a linear axis via a gear box and a ball screw.

![Figure 24. A simplified model of a single axis in a computer controlled machine](image)

According to Bone (1978), the power of a motor increases with the size of its rotor (diameter D and length L) according to,

\[ P = \alpha D^{2.5} L \omega \] (8)
where $\alpha$ is a function of the average flux density in the gap between rotor and stator and the design of the motor windings (Schwarz, 1966, gives more detail on this). The power 2.5 reflects observed performance across a wide range of motors, and accounts for losses from practical electrical, mechanical and thermal loadings. Using (8), the acceleration of the tool in fig. 24 can be calculated from:

$$\alpha D^{2.5} L \omega = J \dot{\omega} + m \dot{v}$$

Therefore, using the standard calculation for the moment of inertia of a solid cylinder (the rotor), and the relation between $v$ and $\omega$ given in fig. 23,

$$\alpha D^{2.5} = \pi \rho \omega L \frac{D^4}{32} + \frac{m \omega}{2 \pi N} \ddot{v}$$

Rearranging, the linear acceleration of the tool is found by,

$$\ddot{v} = \frac{16 \pi \alpha \rho NL D^{2.5}}{\pi^3 \rho L N^2 D^4 + 8 m \rho \omega}$$

Figure 25 now shows how the acceleration of the tool varies with key design parameters. Following common practice, it is assumed that the motor is approximately square, so $L=D$. The set point values are chosen as: ball screw pitch $p=10\text{mm/rev}$; motor constant $\alpha=1$; rotor diameter $D=20\text{cm}$; density of rotor (electrical steel); $\rho=7,800\text{kg/m}^3$; gearbox ratio $N:1=50:1$; tool mass, $m=20\text{kg}$, motor speed (3000 rpm), $\omega=100\pi \text{rad/s}$. Each parameter is increased and decreased by up to a factor of five and the figure shows the resulting tool acceleration normalised by that at the set point.
Figure 25. The sensitivity of tool acceleration to key design parameters of a typical linear axis in machine design.

Figure 25 demonstrates that, for a fixed mass tool, its acceleration increases with ball-screw pitch, and decreases with increasing rotor diameter and gearbox ratio, and this gives important directions for developing faster machines. Relatedly, the tool acceleration will increase with a reduction in the gearbox ratio – ideally to the point of having a direct drive without a gearbox. (The effects of ball screw pitch and gear box ratio are equal and opposite in this analysis as it has been assumed that the gear box has no rotational inertia.) Direct drive of the ball screw depends both on providing a motor with the right torque/speed capacity, but also on the angular precision with which the motor can be positioned which may be a current constraint. However the strongest result in fig. 25 demonstrates a strong benefit in having smaller diameter motors: at the set point data, comparing the inertial terms in eq. (8), the rotational mass is around 40,000 times greater than the linear mass, so reducing the motor diameter is a major opportunity to increase manufacturing speed. Only once the inertia of the tool in (8) exceeds that of the rotor is it worth applying widely known principles of lightweight design to the tool – for example through optimised design and material selection.
The schematic of fig. 24 assumes that the tool must be moved, but does no work – so the analysis should be adapted if a force restraints the tool. Altintas et al. (2011) further provide a detailed review of the dynamics of drive systems spanning a wide combinations of guides and drives and their dynamic performance. However, the principles revealed by fig. 25 suggest a clear benefit in directly driven axes in CNC machines: eliminating the gearbox and increasing the pitch of the ball-screw increases the fraction of the motor’s power that is used to accelerate the tool rather than the rotor. The requirements for gearboxes today appear largely to be to overcome limits in the precision with which the rotor shaft can achieve a specific angle. Eventually, as the rotor diameter is reduced, it will reach the minimum size with sufficient torque to provide a required acceleration, at which point the rotor length $L$ should be extended preferentially over its diameter. This length is constrained by vibrations of the rotor as it rotates, but the series connection of several narrow diameter rotors would remain preferable to increasing rotor diameter – due to the fourth power scaling of the rotor’s moment of inertia with diameter against its linear dependence on length.

3.3 Control of heat transfer

Figure 26 illustrates a simplified heating problem representative of many manufacturing processes: a workpiece, initially at uniform temperature $T_0$, and with its smallest dimension $2X$, is to be heated by contact with a heating medium (which might be air, a fluid, or a solid) at constant temperature $T_m$. The centre of the workpiece at $x=0$, must be raised to a temperature $T_f$ as rapidly as possible, while the surface temperature (and hence that of the heating medium) must not exceed some value $T_{lim}>T_0$ to avoid melting, or other unwanted material transformations.
The evolution of temperature in the workpiece is described by,

$$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2}$$

subject to $T(0,x) = T_0, \quad 0 \leq x \leq X$

$$\frac{\partial T}{\partial x} \bigg|_{x=0} = 0$$

and

$$k \frac{\partial T}{\partial x} \bigg|_{x=X} = h_c \left( T_m - T(X,t) \right)$$

and $T(X,t) \leq T_{\text{lim}}$

where $\alpha$ is the thermal diffusivity of the workpiece, $k$ its thermal conductivity, and $h_c$ is the thermal contact conductance of the interface between the workpiece and the heating medium. A typical range of the key parameters of eq. (12) is selected for illustrative purposes and given in table 2, assuming that the heating medium is solid, and accounting for the potential increased thermal contact conductance that occurs with increasing interfacial pressure between the medium and workpiece. The workpiece is made of aluminium, in this example.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Low value</th>
<th>High value</th>
<th>Set-point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workpiece half-width, $X$, (mm)</td>
<td>2.5 (sheet)</td>
<td>250 (forging)</td>
<td>25 (plate)</td>
</tr>
<tr>
<td>Temperature excess for $T_m$ over $T_r$ (K)</td>
<td>2.5</td>
<td>25</td>
<td>250 (water)</td>
</tr>
<tr>
<td>Thermal contact conductance (W/m$^2$K)</td>
<td>100 (paper on metal)</td>
<td>1000 (metal on metal, low pressure)</td>
<td>10,000 (metal on metal, high pressure)</td>
</tr>
<tr>
<td>Initial workpiece temperature $T_0$ (K)</td>
<td></td>
<td></td>
<td>293</td>
</tr>
<tr>
<td>Final workpiece temperature $T_f$ (K)</td>
<td></td>
<td></td>
<td>393</td>
</tr>
<tr>
<td>Workpiece thermal conductivity $k$ (W/mK)</td>
<td>237 (aluminium)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workpiece specific heat capacity $C_p$ (J/kgK)</td>
<td>900 (aluminium)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workpiece density $\rho$ (kg/m$^3$)</td>
<td>2740 (aluminium)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Typical process parameters for heat transfer between a workpiece and a solid as might occur in manufacturing

Given a selection of values from table 2, eqn. (12) can be solved numerically, to find the time required for the centre-line workpiece temperature to reach $T_f$. This is done firstly for the set-point values, and then with each parameter varying in turn over the range specified, and the normalised results are shown in fig. 27.
Figure 27. The sensitivity of heating time, for the centre of a workpiece, to reach a set temperature in response to variation in over-heating, the interface thermal conductance and the smallest workpiece dimension.

Figure 27 shows a strong connection between the speed of heating to the required centre-line temperature, and all three parameters: the heating time falls as the temperature excess and thermal conductance increase and as the workpiece minimum dimension reduces. In many cases – as discussed above for the sintering of nylon 12 powder – the allowable excess temperature is tightly constrained by material limits, so this excess must avoid any transition to other material behaviour. However the other two curves of fig. 27 give key insights into opportunities to double manufacturing speed:

- Heating time will halve if the thermal contact conductance is doubled. Although the data here assumed solid-to-solid contact, this conductance increases as the heating medium moves from gas to liquid to solid (although the model would require extension to include convection to incorporate this), and within the solids, increases with the conductivity of the material and the contact pressure.
- Heating time, in this example, will also halve if the minimum dimension of the workpiece is halved.
Revisiting the discussion of section 2 with these insights suggests some interesting opportunities for process development. The current speed of laser sintering of polymer powder is constrained by the time required to pre-heat the powder bed, and to allow it to cool after sintering is complete. However, the polymer powder itself has poor thermal conductivity: could a different powder, for example a fine metallic powder with much higher thermal conductivity, be used as a filler in areas where no sintering will occur? Is it possible to apply pressure to the powder bed to increase its conductivity? Similarly, the discussion on hot rolling of strip discussed the need for simultaneous control of temperature, strain and strain rate to achieve specified grain sizes. Could the strip be cast thinner to allow faster adjustment of strip temperature? In bulk forming processes such as forgings, where components are typically thicker than plate, is it possible to apply the high temperature forming processes to nearer to net shape workpieces – to allow more rapid and precise control of temperature and hence grain size right through the workpiece?

3.4 Fluid flow

Fluids are used in many processes at the interfaces between moving components, for lubrication or cooling, and also for general temperature regulation. A key constraint on manufacturing speed is the rate at which fluid can be delivered to the required location (the volume flow rate), and a common mechanism to achieve this is to pump the liquid along a circular pipe of diameter $D$. For a fluid with density $\rho$, and viscosity $\mu$ travelling at an average speed of $v_{\text{avg}}$, the flow is characterised by its Reynolds number,

$$\text{Re} = \frac{\rho v_{\text{avg}} D}{\mu}$$

(13)

For Reynolds numbers below $\sim 2300$, the flow is laminar, and for fully developed flow (that is, when the flow has travelled along a sufficiently long pipe that the speed of the fluid is independent of its axial location), the volume flow rate is characterised by Poiseuille’s Law (see for example Çengel and Cimbala, 2006) as,

$$\dot{V} = \frac{\Delta P \pi D^4}{128 \mu L} = \frac{h_i \rho g \pi D^4}{128 \mu L}$$

(14)
Here $\Delta P$ is the pressure drop driving the flow along a length of pipe, $L$, which can equally be expressed in terms of the head loss $h_L$ where $\Delta P = h_L \rho g$. For turbulent flow, the volume flow rate depends on friction between the pipe and the fluid, which in turn depends on the surface roughness of the pipe, characterised by its mean height, $\varepsilon$. This friction factor is represented on a Moody chart, and can be used to find the volume flow rate by iterative solution of an implicit equation. However, to simplify this process, Swamee and Jain (1976) proposed an explicit approximation that for Reynolds numbers greater than about 2000, the volume flow rate is,

$$\dot{V} = -0.965 \left( \frac{gD^5 h_L}{L} \right)^{0.5} \ln \left[ \frac{\varepsilon}{3.7D} + \left( \frac{3.17 \mu^2 L}{\rho^2 gD^3 h_L} \right)^{0.5} \right]$$  \( (15) \)

Table 3 presents typical ranges of the parameters that might apply when pumping fluids along circular pipes in manufacturing processes, with a nominal ‘set point’. This allows construction of fig. 28, based on equations (13) and (15) for laminar and turbulent flow respectively, which shows how the volume flow rate varies from the set point ($\dot{V}^*$) as each parameter is varied in turn.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Low value</th>
<th>High value</th>
<th>Set-point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluid density, $\rho$, (kg/m$^3$)</td>
<td>785 (acetone)</td>
<td>1260 (glycerol)</td>
<td>1000 (water)</td>
</tr>
<tr>
<td>Viscosity, $\mu$, (mPa s)</td>
<td>0.3 (acetone)</td>
<td>1500 (glycerol)</td>
<td>1 (water)</td>
</tr>
<tr>
<td>Pipe diameter, $D$, (mm)</td>
<td>2.5</td>
<td>250</td>
<td>25</td>
</tr>
<tr>
<td>Head loss, $h_L$, (m)</td>
<td>0.1</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Pipe Length, $L$, (m)</td>
<td>0.1</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Roughness, $\varepsilon$, (mm)</td>
<td>0 (glass)</td>
<td>0.01 (rubber)</td>
<td>0.0015 (copper)</td>
</tr>
</tbody>
</table>

Table 3. Typical process parameters for fluid flow in circular pipes as might occur in manufacturing processes.
Figure 28. The sensitivity of volume flow rate to variation in key design parameters, relative to the nominal set point shown in Table 3.

Figure 28 (in which almost all the results are in the turbulent flow regime) demonstrates that the key to increasing fluid flow is to increase the diameter of the pipe. The head loss and pipe length have inverse effects, as the head loss per unit length is the key metric, and is the next most influential parameter. The fluid type and pipe roughness have negligible effect.

The results of fig. 28 are illustrative – many other fluid flow problems could have been analysed similarly, and for example, an extension to this analysis would demonstrate the importance of avoiding sharp bends in fluid pipes. However, the key focus on pipe diameter and the relative insignificance of fluid selection, gives clear guidance for applications where the fluid flow is used for thermal stability. In applications, such as in grinding, where the value of the fluid depends on its proximity to the interface, the result suggests that flooding the interface with liquid – perhaps submerging the entire operation in liquid – might be a more effective mechanism to achieve lubrication than adjusting the composition of the liquid. In cutting where the critical ‘pipe’ is the gap between the tool and the chip, a reduction in viscosity is beneficial.
4. Speed limits for manufacturing systems and co-ordination

In parallel with the constraints on physical process operation discussed in the previous section, section 2 also revealed three features of the wider system of manufacturing that may impose fundamental constraints on speed, and these are examined in this section.

4.1 Co-ordination of resources in the face of variability

Several aspects of the manufacturing system depend on co-ordination between different tasks – whether in the sequencing of work through a job-shop, in logistics decisions with supply chain partners, or in managing the complex flow of information in a new product design process. The root of the difficulty in speeding up this co-ordination lies in the variability between the time required to complete different connected tasks, and the uncertainty about how long each task will actually take.

To examine this problem of co-ordination, fig. 29 proposes a simplified model which might be a job shop or a form of Gantt chart for a project. The model has a number of processes, which produce a set of products. A product is defined by the sequence of processes through which it must pass, so in the simplest case each product would require one process only, while in the worst case, each product would have a unique route through every available process. The duration of each process is characterised by a normal distribution, with fixed mean and variable standard deviation.
Figure 29. Simplified simulation of a job shop to explore resource co-ordination.

The simulation is run (using a discrete event model) with a large number of products (1000), to allow long-run behaviour to emerge. Each run is set up with a choice about product variability (characterised by the number of unique product designs), the number of processes (between 1 and 8), a fixed mean time for all processes but variable standard deviation of the process duration (calibrated so that three standard deviations is set to be between 0 and 75% of the mean). Figure 30 shows how the time taken to complete each of the first 200 products in the series increases from the ideal minimum as each of the three variables in the system is varied from its mean value. (In order to clarify the trends of these plots, each simulation was run 100 times, with different randomly generated products and process durations, and every line on these graphs is thus the average of these 100 runs.)
Figure 30. Performance of the system of fig. 29 in the face of increasing product variability, number of processes and process duration variability.

The colours in each plot indicate the number of processes in use; moving from top to bottom of the grid of plots, the processing times become increasingly variable (the standard deviation of process times increases with alpha); moving from left to right of the grid of plots, the number of different products increases, beta indicating the fraction of all possible process routes in use.

The top left plot of fig. 30 shows that, as expected, if there is no variability in process times, and all products are identical (require an identical sequence of processing), then the process time remains at the idea level for all production. This is the ideal target pursued by Henry Ford, and more recently in the Toyota Production System: product variety is introduced in such a way that processing times at each process station are unaffected, and variety is driven out of the system by drawing attention to any deviation from expected behaviour.

Moving down the left column of fig. 30, beta remains zero indicating that production remains perfectly serial, but the process time variability increases –
and regardless of the number of processes, the time to complete each product rises as production continues. This is because process time variability leads to the build up of queues in front of each process, and eventually the queues reach a sufficient length to ensure that the process always remains busy, and the queue ceases to grow longer. The variability is largely unaffected by the number of processes in the line, as once this queue length is established for one process, it will reach similar levels for the other processes, but on average, no individual product will have to wait for the full queue duration at all processes. This result repeats for all other columns of the figure – in every case, increasing variability in process time affects the single process design as much as the system with many processes.

In contrast, reading across the rows of the figure, adding variety to the product mix – reflected by having different routes through the production system – instantly increases the production time per part dramatically, as queues build up in front of processes due to conflicting demands rather than variability. For the systems with four or eight processes, the influence of product variety occurs as soon as beta (the metric of product variety) moves from zero, and increasing variety further has little further effect – once conflict has been created by products moving in different directions through the system, queues build up in front of processes, and these reach a sufficient level that further increases in product variety have little further effect. Any set-up or switchover time for a process changing from one product to another would dramatically increase this effect.

The implications of fig. 30 are striking – not just for the co-ordination and scheduling of work in a job-shop, but for any task such as design, cost management or logistics, requiring the co-ordination of multiple resources across several tasks. Once the same resources are involved in more than one overall chain of activity, queues of waiting work build up, and the time to complete any single task increases. This clearly supports widespread work to group activities into families that have similar characteristics – to facilitate the flow and co-ordination of activity among different processes. Once this grouping has been achieved, the figure shows that efforts to reduce the variability of task
times will be rewarded with reductions in overall task completion time – as queues of waiting work can be eliminated. This result applies as much to the co-ordination of office work or service systems as it does to the mechanical effects of production.

4.2 Biophysical limits to human performance

Two aspects of the biophysical limits of human performance influence the potential for increasing the speed of manufacturing operations: the avoidance of musculoskeletal disorders from task repetition; the design of tasks to make optimum use of human biomechanics. The first of these areas has received significant attention within literature associated with manufacturing, while the second is as yet largely an area of scientific research, with applications mainly in the areas of sports science and rehabilitation.

The computerised Ovako Working Posture Analysis System (OWAS) (introduced by Karhu et al., 1977) has been used widely for characterising tasks in manufacturing. For example, Mattila et al. (1993) use the system to examine the origin of musculoskeletal disorders in 18 construction workers doing hammering work and relate them to particular postures, while Balogh et al. (2006) use detailed observations in a factory producing parquet flooring to relate musculoskeletal disorders to particular combinations of activity.

The OWAS system allows a structured analysis of existing work practices, but to be useful for design of new tasks, it must be codified in some form of model to anticipate postural problems in future. In pursuit of this objective, Kingma et al. (1996) use ‘linked segment models’ to characterise body motions and validate their model through detailed experiments on the ground forces measured in practice as human subjects lift known weights across various carefully selected trajectories. Damsgaard et al. (2006) describe the use of an “Any Body” modelling system, they have developed to attempt to mimic the rigid-body motions of the human musculoskeletal system, parameterised to recreate “any” body. They pose the “Muscle recruitment problem” as an optimisation problem, and show how humans choose particular muscles to complete some specified task by minimising muscle forces.
The outcome of work on postural analysis is typically a recommendation for some sort of task redesign; it appears that postural problems can generally be solved, for example through different work positioning or through the use of lifting devices, so do not present a limit on future manufacturing speed. An opportunity for future development might be the exploration of whether manufacturing tasks that require manual labour might be redesigned to allow more efficient use of the working potential of human muscles. A rich literature has been triggered by the first analysis of muscle dynamics (Hill, 1938) that seeks both to characterise the delivery of power by muscles, and to optimise the application of power in particular applications. For example, the experimental tests on an insect muscle by Josephson (1985) or those by Ahn et al. (2003) on a hopping toad, demonstrate that the work-loop created by a muscle (the loop created by combining the force against time and extension against time plots of the muscle during one full cycle of extension and contraction) depends strongly on the timing of the stimulus. The total work done by the muscle is maximised only if it is stimulated to act at the correct point in the cycle. Potentially, this form of muscle characterisation could in future be combined with models of human movement (for example as reviewed by Pandy, 2001) and requirements for muscular effort (reviewed by Erdemir et al., 2007) to find tailor the design of tasks to optimise the capacity of human bodies and thus contribute to future manufacturing speed increases.

4.3 Cognitive limits to human performance

Despite the extraordinary adaptability of human beings, the speed of their cognitive processes has limits, and in turn this will provide a limit to the absolute speed at which manufacturing can occur: despite extensive ‘hype’ about robots taking over human jobs, the reality of the past 40 years of automation is that robots or other computer operated systems are less able than humans to respond to unfamiliar situations, so while highly controlled and repeatable tasks can be automated, it is unlikely that manufacturing will ever be independent of human control, support, innovation, leadership and repair. The fact that such limits exist is recorded in several studies which detect physical signals of internal cognitive limits. For example, May et al. (1990) show that the range of
spontaneous eye movements (detected with an infra-red eye-tracking instrument) decreases as cognitive workload increases indicating an increased focus on more specific tasks. Ahlstron and Friedman-Berg (2006) also examined eye movement in air-traffic controllers during adverse weather conditions, with or without a dynamic storm forecast screen, finding a correlation between higher cognitive workloads and shorter blink duration and pupil dilation.

However, a broader issue is not so much about the absolute limits of human cognitive performance as how humans in work make decisions as their job characteristics vary. This is the subject of “cognitive work analysis” a rich field of study, largely referenced to key works by Rasmussen and colleagues in Denmark, such as Rasmussen et al. (1994). This approach to analysing work systems has been described by Vicente (1999) as having five phases:

- Work domain analysis leading to a description of the functional structure of the work being considered, including the purposes, priorities, functions, physical processes and physical objects involved.
- Control task analysis describing the interventions that allow work to complete, including roles and decision ladders.
- Strategies analysis identifying the ways that each intervention might be carried out.
- Social organisation and co-operation analysis of how people and computers interact during the work.
- Worker competencies analysis to identify the knowledge, rules and skills needed by each worker in the system.

The intention of this analysis is to inform the design of work, spanning information flows and representation, task specification, worker training and skills requirements, and the social structures of work teams, to improve overall efficiency.

Hassall and Sanderson (2014) focus on the ‘strategies analysis’ phases of this sequence, examining both the range of strategies that workers might consider in addressing a particular task, and the factors that shape their selection among the various strategies. They represent their analysis using the ‘strategies cube’
shown in fig. 31: the cube shows eight strategies organised on three axes of constraints: time pressure, risk level (the degree to which the system or its components are exposed to harmful or unwanted outcomes) and difficulty in execution. A task with high difficulty, but subject to low time pressure and risk, might be addressed with analytical reasoning, where a similar task addressed under high time pressure and with high risk might instead lead to the undesirable strategy of ‘arbitrary choice.’

**Fig. 31. The strategy cube proposed by Hassall and Sanderson (2014).**

While not intended to address the question posed in this paper, Hassall and Sanderson’s analysis provides a framework in which it might be considered: the organisation of strategies in the cube suggests that as manufacturing speed increases (and so time pressure increases), workers will tend towards the strategies on the front face of the cube. As “Avoidance” and “Arbitrary choice” are obviously undesirable, this suggests that increasing manufacturing speed requires a reduction in the difficulty in executing tasks. For low risk tasks, the “option-based” and “imitation” strategies, assume that the tasks are sufficiently standardised that workers can respond either by selecting among pre-defined options, or by imitating a standard procedure. Naïve compliance with standards was both the inspiration of Ford’s production system, and its limitation (the
system responded poorly to deviations from the expected standards), which has largely been superseded by Toyota’s Production System (which is specifically designed to draw attention to deviations from standards to allow preventative action and build greater understanding into the system.) The implications of this cube then are that if the complexity of a full manufacturing system is to be operated at double the speed, the tasks to be performed by the workers in the system – at all levels, from production to design, accounting and management – should be designed in a way that creates as much similarity as possible between tasks in different contexts, to allow imitative and option-based strategies to develop and be refined.

5 Promising current developments

Sections 3 and 4 have examined some key limits to manufacturing speed: clear understanding of fundamental limits to speed gives a basis for identifying new opportunities to pursue faster manufacturing in future. In parallel with this more theoretical analysis, this section reviews five current and promising developments in the hope that an approach to speed improvement in one area may be transferable elsewhere.

5.1 Moving from Point to Line to Array scanning

Several of the processes reviewed earlier in this paper rely on scanning a point actuator or sensor across a workpiece. This was the case with incremental sheet forming, laser sintering of powder and tactile metrology for surface topography for example. An obvious opportunity to increase manufacturing speed is to replicate this single point – in a line of parallel actuators or sensors, or most richly, in a scanning array. This approach is under development in several areas, but seems to be most developed in inkjet printing.

The simplest inkjet printhead consists of a single nozzle from which drops are ejected in response to a data stream. In order to produce a pattern of dots on a two-dimensional substrate, the substrate and the nozzle (or the point of impact of the drops) must be moved relative to each other so that each discrete location on the substrate can be addressed. This is illustrated in fig. 32.
Figure 32. Schematic diagram showing how, in principle, production speeds in inkjet printing can be increased by changing from single-nozzle scanning, to multi-pass and single-pass printing, and then by the use of a multiple nozzle, two-dimensional array printhead.

Serial scanning with a single nozzle requires a time $T$ for the nozzle to pass over every location. By using a printhead with multiple nozzles and scanning it back and forth across the substrate the processing time can be much reduced to $T/n$ for a printhead with $n$ nozzles, ignoring any delays due to the change of direction. This ‘multi-pass’ method is commonly used in commercial inkjet printing and in office printers and there has been a steady increase in the numbers of nozzles in printheads which has led to increased printing speeds. The next stage of development is to use a printhead which can print across the whole width of the substrate at once, and to move the substrate continuously beneath the printhead.

For a printhead with $N$ nozzles the time is now $T/N$. This increase in printing speed is very attractive, but there are disadvantages as well. In ‘single-pass’ printing with an array, the failure of a single nozzle could lead to unacceptable loss of quality in the product, for example through an unprinted line on a graphical image, or a critical defect in an electrical conductor. While it is possible to compensate for defective nozzles in various ways, the single-pass method tends to be less forgiving than a multi-pass method in which the printhead moves across each part of the substrate more than once. Nevertheless, the single-pass process is being successfully employed for high-speed applications such as label printing and high volume office printing. Multi-pass and single-pass methods represent the current state of the art; in principle, at least, one might
consider the use of a two-dimensional array of nozzles which would reduce the printing time still further (to $T/N^2$ in the illustration in fig. 31), but time would be needed to re-locate the substrate after each printing event and there would be major challenges in designing and constructing the printhead itself. It seems likely that single-pass printing on to increasingly wide, continually moving substrates, possibly with more than one printhead in tandem, will prove to be the best method to increase the production rate. Certainly, there is a current commercial trend to increase the number of nozzles in a single printhead.

An interesting application of multiple-nozzle inkjet printing to additive manufacturing, described by Ellis et al. (2014), is the High Speed Sintering process. To overcome the speed limits of the single laser spot actuation of conventional laser sintering, a pattern of radiation absorbing material is inkjet printed over a whole layer of the thermoplastic powder bed, allowing simultaneous sintering of the whole layer by irradiation from an infrared lamp.

The same principle has been applied to the incremental sheet forming process discussed in section 2.2.2 by Kwiatkowski and Tekkaya (2015). They propose and demonstrate the extension of single point incremental sheet forming to two, four or an array of point tools. By exact analogy of the development illustrated in fig. 32 above, their intention is to reduce processing time through applying multiple point tools in parallel.

5.2 Parallelisation

A related approach to this development of line-scanning is being developed for material removal processes: parallel or simultaneous machining employs more than one cutting tool on a workpiece. It offers potential increase in material removal rate/productivity due to increased number of tools. This idea has been applied in turning and milling and named as parallel turning and parallel milling, respectively. Parallel grinding and drilling applications are also under development.

Early research in this area has explored the advantages of parallel turning (fig. 33a) and parallel milling (fig. 33b). In parallel turning, the workpiece surface cut by one tool meets with the other tool half a revolution later. Hence, the
vibrations created by one tool will interact with the vibrations created by the other one. In some cases, this leads to considerable benefit in reducing overall workpiece vibration (Budak and Ozturk, 2011) and if the two tools are diametrically opposed and perfectly controlled, the bending force applied to the workpiece is reduced to zero creating an additional improvement in accuracy.

Figure 33. (a) Illustration of the parallel turning process (b) Photo from a parallel turning operation (Budak and Ozturk, 2011)

In parallel milling, illustrated in fig. 34, two tools have been applied on a flexible workpiece. As with parallel turning, there is a dynamic interaction between the tools. Process models which have been developed to understand this interaction, for example by (Ozturk and Budak, 2010) can be used to maximize this benefit, with experimental results showing 100% improvements in productivity.

Figure 34. (a) Illustration of the parallel milling process (from Budak et al., 2013) (b) Parallel milling experimental set-up at AMRC

As parallel machining technology is developed, either new machine tools with new kinematics will be needed, or robotic machining technology can be
developed, subject to the well known trade-off between the benefits of robot flexibility and their increased dynamic flexibility.

5.3 Tailored application of precision

A different approach to increasing the speed of scanning processes is to modify the degree of precision with which actuation or sensing is to be applied, according to either customer specification or the errors predicted by a process model. This approach depends on exploiting a priori knowledge about the production task, the nature and functional significance of relevant defects, and potential repair steps in order to simplify the task.

This tailored approach to precision is being developed with particular attention in the area of metrology, which requires new insights into: optical system modelling and defect extraction, global control of substrate and defect location, intelligent sampling and data throughput, and fast feature inspection and multi-scale calibration support. The measured surface data must be simplified, using standardised methods and filters, to retain a manageable set of representative parameters or dimensions (with associated uncertainty information) to describe and distinguish defects to correlate with function (Bruzzone et al., 2008). Specification and standards for calibrating geometry and surface topography measuring instruments are under current development (Wang et al., 2015).

If defects across the surface of interest are correlated and have a cumulative effect, representative areas may yield enough information for inline decision-making; an example would be tool-wear effects on embossed lens arrays. Alternatively, if a single local defect can have a large functional effect, such as for breaks in printed conductors, 100 % inspection might be required, but could be constrained to pre-defined bands.

One potential solution to implementing tailored precision is the use of hybrid instrumentation, i.e. to detect areas of interest with a relatively low resolution sensor (for example, camera-based sensors), then to “home-in” on the areas of interest using a localised, high resolution sensor (see for example, Kayser et al., 2004). In some cases the low resolution sensor could use approaches that detect
light scattered from points of interest, such as defects or scratches, therefore, allowing high resolution detection without the need for imaging.

An alternative solution to the requirement for precision is used in micro-chip manufacture. In this case, testing the batch of chips on a wafer can be achieved relatively rapidly and it may be cheaper to allow production of a batch with some known probability of faults and then to find and cut them out, than to increase precision to guarantee their elimination. This approach to speed increase is discussed for example by Huang et al. (2003) and is described elsewhere as “fault-tolerant” batch production.

Where two or more sensors are used with different lateral resolutions, data fusion techniques are required to combine the data and to match the co-ordinate systems of the sensors (Weckenmann et al., 2009). To reduce scanning times and make full use of all the available information, intelligent sampling techniques must be used (Wang et al., 2012), such as compressed sensing (Baranuik 2011), although this has yet to be demonstrated in practice.

Metrology is often the bottleneck in a manufacturing process. In many case, the conventional use of metrology techniques is reaching its limits: speeds might be increased slightly by employing faster processors or scanning mechanisms, but several orders of magnitude speed increase are required to keep up with modern, highly-parallel manufacturing techniques. To address this, new ways to use existing technologies are required. Leach et al. (2014) survey some conceptual approaches including measuring the difference from nominal topography (i.e. defects) rather than the actual topography and using a priori manufacturing data (e.g. from CAD systems) to increase the spatial bandwidth of the measurement process while reducing the time required.

5.4 Hybridisation and supply chain compression

Sections 5.1 to 5.3 have discussed three approaches to increasing manufacturing speed where the constraint is the time required to move a point actuator or sensor. A different opportunity related to the speed of the whole manufacturing system is to exploit opportunities to compress the length of an existing supply chain through process hybridisation. This approach has been reviewed across a
wide range of applications by Lauwers et al. (2014), and a representative example of this approach is the development of laser assisted Jet ECM (LAJECM), in which a laser beam is directed through the centre of the ECM tool at the active electrolyte. The heating created by this additional energy source has been shown to improve dimensional accuracy by 78%, and increase production rates by 46% (De Silva et al., 2003). Figure 35 shows a comparison of volumetric removal rates in LAJECM and non-laser JECM for Hastelloy (Pajak et al., 2006).

Figure 35. Comparison of ECM performance with and without laser assist

A different example of hybridisation is the development of fixtures with intelligent embedded sensing that may be tuned specifically to the part being processed (Papastathis et al., 2012), to collect extra process data such as geometrical uncertainties or systemic dynamic behaviour. This information can then be fed-back into the manufacturing system to allow in-process corrective action (Abellan-Nebot et al., 2012). Intelligently designed fixturing with embedded actuation can also be used to tune a manufacturing system’s dynamic behaviour, enabling improvements in both material removal rates and surface finishes (Kolluru and Axinte, 2014). Smart and active fixturing can increase production through-put via improved system stability.
5.5 Design taxonomy and beyond

A feature of several of the system and co-ordination functions discussed in section 3 was that the constraint on increasing manufacturing speed related to the difficulty of keeping several different resources (machines, people, supply chain partners) working efficiently, due to the combination of uncertainties in how long each task would take. A well trodden approach to addressing this problem is to attempt to specialise every task in the supply chain, so it is better understood, more repeatable, and by being easier to predict, is easier to co-ordinate. However, this approach is constrained by the fact that across the wide variety of products in manufacture, and the supporting functions that allow innovation, there is a limit to how much standardisation can be achieved. An alternative approach being developed in design, and with potential applicability across a wider range of other manufacturing functions, is to develop taxonomies of previous tasks.

Atman et al. (1999) conducted a series of longitudinal studies on a number of student classes, and noted that in order to create efficient, better designs quickly, it is important that the problem-definition and requirements list is correctly defined in the first instance. However, he also pointed out that this is typically related to experience and that a novice designer who gathers considerable quantities of data during the scoping phase may not translate to better designs in the concept design phase. Conversely the expert designer tended to gather more information, created more wide ranging alternatives and transitioned frequently between the different design phases. Ahmed et al. (2003) also noted that the novice designer will tackle design in a trial and error manor, therefore increasing design time, however the experienced designer tended to use more integrated design strategies that involved making critical design based judgments before fully executing an idea, thereby reducing the design time. Whilst these research studies depict the value of expert knowledge, capturing this knowledge particularly for new product design can be challenging.

From a purely mechanistic viewpoint, increased design speeds have been achieved to some degree with the application of model based engineering techniques. These have taken the form of product taxonomies, used to define the
predominant characteristics of a product family; classifications to define particular facets of a product and parametric design rules applied to a product family, for providing a fast method with which to customise a design. Parametric techniques are often used to provide a method for designers to change a design rapidly based on a set of predefined rules. For example, this can be in the form of mathematical expressions that allow alterations in the size and shape of products. This approach allows for re-use of existing products and rapid design modifications based on engineering analysis (Myung and Han, 2001).

6. Discussion

The introduction to this paper recognised that doubling the speed of manufacturing as a whole may require simultaneous increases in the physical processes of production, the system which connects them, and in the wider co-ordination of the supply chain, design and accounting functions for example, which provide the context of manufacturing. The argument of the paper has been to reveal that the constraints to speed increase in several contrasting case studies can be reduced to a small number of common issues, and a simplified analysis of each issue has given some insight into the absolute limits to speed improvement. Contrasting these limits with current practice suggests some opportunities for innovation, and section 5 has reported on some current developments promising a step change in the speed of particular functions of the manufacturing system; there are clearly significant opportunities to speed up production.

Not all opportunities to increase manufacturing speed will prove attractive in practice. Even if manufacturing at double the speed is possible – with sufficient innovations in processes, system and co-ordination – this still may not be a desirable goal. The design of manufacturing requires a balancing of many objectives, including operational and investment costs, the overall size and rate of change of a market, and the human goals of those involved in production and their customers.
The introduction to the paper made a specific exclusion to its scope about the development of new materials. In practice, many of the processes reviewed in section 2.1 are currently constrained either by tool or workpiece material properties, so material development is currently a key part of the effort to increase speed in manufacturing. For example, solid polycrystalline diamond-tipped drills have recently been developed to overcome physical limits to drilling composite materials through increased tool life, quality and productivity. In turn, development of a new material that solves one constraint on manufacturing speed, may simply shift attention to a different constraint. For example, higher speed machining of exotic alloys has been enabled by the use of high hardness tool materials such as CBN or ceramics, but these are sensitive to vibration so the constraint on speed changes from tool wear to the dynamic stability of the tool and workpiece assembly and the fracture toughness of the tool material. Relatedly, with the advent of new materials for tools or workpieces, process modelling may become a constraint on speed, until new material characterisation is achieved.

Furthermore, new processes may themselves enable the development of new manufacturing equipment that allows increased speed. "Rapid tooling" was one of the early motivations for the development of additive manufacturing processes, for example to the incorporation of cooling features within more complex tooling structures. However, there are many further opportunities (some well explored and some proposed) which will also create enhanced functionality tools. These may for example result from the incorporation of harder materials within a single tool component, surface engineering for enhanced tribological properties or reconfigurable tooling. Tooling created by various additive manufacturing processes has been proposed for forming operations, for example Hölker et al. (2012) propose the production of hot extrusion dies by laminating laser cut layers and Hölker et al. (2013) similarly make dies by selective laser melting. However the approach may also be applied much more broadly. For example Arthur et al. (1996) explore the development of EDM electrodes by additive processes.
Nevertheless, the objective of this paper, which was to pose a hypothetical question, and explore the constraints on existing practice that prevent its immediate solution, has led to several interesting suggestions for innovation. Without doubt, the question addressed in this paper could be examined in greater depth and by a wider community to enrich and expand the suggestions made here. The processes considered in section 2 were the subset known to the paper's authors, and wider examination of other processes would enhance the paper's coverage. Furthermore, other similar questions could stimulate related searches for unexpected innovation: is it possible to manufacture with half the energy? Can we design and make products that remain valuable to customers for ten times longer? If we only had three machines to manufacture a lifetime's products, what would they be?

Manufacturing at double the speed is certainly possible in many cases, and the pursuit of speed seems likely to reveal other opportunities for valuable innovation.

Acknowledgements

The first author's work on this paper was funded by EPSRC Grant EP/K018108/1 and Professor Hutchings's by EPSRC Grant EP/H018913/1.
References


in ergonomics. *International conference on Digital Human Modelling, San Diego, CA, USA, 591-598.*


Muller, J.-D., Bousmina, M. and Maazooz, A., 2008. 2D-sintering kinetics of two model fluids as drops, *Macromolecules 41, 2096-103*


