Computing for the Future of the Planet

By Andy Hopper and Andrew Rice

University of Cambridge, Computer Laboratory, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK

Digital technology is becoming an indispensable and crucial component of our lives, society, and the environment. We present a framework for computing in the context of problems facing the planet. The framework has a number of goals: an optimal digital infrastructure, sensing and optimising with a global world model, reliably predicting and reacting to our environment, and providing digital alternatives to physical activities.

This paper describes our vision in which datacentres can scale power consumption in line with performance, run closer to the wire with reduced redundancy, and behave as a "virtual battery" dynamically utilising spare, or otherwise unusable, generation capacity from renewable sources. On a broader scale we consider how global sensing might allow us to optimise our daily activities and lives. We highlight the issues and dilemmas inherent in the deployment of global sensing infrastructure and we work towards our challenge of a Personal Energy Meter as a tool for informing decisions and providing impetus for reducing the ecological footprint of our society.

Keywords: computing, environment, energy, optimisation, sensing, datacentre, sustainability

1. Introduction

Computing's great success has been its use as a general tool: from scientific computation on the first computers 60 years ago, to the billions of computing devices pervading business and society today. Computing is now undergoing a shift from use as a general tool to existing as a fundamental resource built into our environment. Commonplace activities, such as Internet web search, mobilise huge computing resources on our behalf. The size and scope of these activities would have seemed unobtainable a mere 10 years ago and yet we take this for granted, and even depend upon it, without realising that we do so. Computing is becoming mandatory for our day to day lives.

Computing for the Future of the Planet is a framework for identifying ways to develop this technological resource so that computing can have a positive effect on our lives and the world (Hopper 2007). This framework is useful in identifying problems to tackle with technology, for defining the scope of these problems and highlighting the potential impact of their solution. We believe that computing could provide alternatives to our current activities and, through availability of information and education, an impetus for changing our lifestyles. One might argue that a total shift from physical to digital seems unlikely in today's world but for future generations this concept might seem as obvious as email is to us today.

In Phil. Trans. R. Soc. A 366(1881):3685–3697, 2008
Please see http://journals.royalsociety.org/content/c5241623l1g077nn/

We have identified four principal goals within the Computing for the Future of the Planet framework. Our first goal considers an *optimal digital infrastructure*. Koomey (2006) estimates the energy consumption of computing in the US to be growing at a rate of 15% per year and to consume approximately 1.2% of total energy demand. This figure is conservative as it only considers the energy cost of using and cooling servers and omits network infrastructure and data storage systems. Furthermore, it is important to consider the entire lifecycle cost of the machine. Williams (2004) finds that the manufacturing phase of a desktop machine with CRT monitor consumes an estimated 6400 MJ, which is 35% of the energy used in its entire lifetime, if one assumes a typical case of three years of continuous operation. An optimal digital infrastructure maximises the potential environmental benefits of computing by making efficient use of the energy consumed in manufacture, operation, and disposal.

Secondly, we wish to *sense and optimise* the world around us with reference to a global world model. This model will inform us about the energy consumption and other effects of our activities on the natural environment. Efforts to reduce our ecological footprint will require detailed information to evaluate their success. Furthermore, through optimisation, sensor information potentially allows us to minimise the energy consumption and footprint of our physical infrastructure. Sensing the planet requires a global network of data sources and means for sharing and trading data within it.

Our third goal is to *predict and react* to future events in the natural systems around us by modelling their behaviour. Scientists continue to strive to improve the accuracy of these models by refining the physical models and incorporating new systems within them. Computing has a role to play in helping provide guarantees about the correctness of these models. This considers both the accuracy of the concept with which we model a physical process and errors embodied in the implementation of this concept. The accuracy of the physical-process aspect of a model might be examined with visualisation tools that monitor the evolution of the modelled state. This may help identify shortcomings in models and decide the required level of modelling fidelity. Implementation reliability might be approached using the same theoretical correctness techniques that have already been used to improve the security of software. The traditional role of computing as an execution platform for these models will continue to be important and must grow in performance to service both the increasing demands of higher-fidelity models and also to accommodate any new overheads incurred by correctness checking.

Finally, we are interested in the possible benefit of *digital alternatives to our physical activities*. Examples of this include electronic versions of printed newspapers, music downloads and digital music players rather than physical CDs, and online shopping as opposed to visiting the high-street or supermarket. It is not clear whether the impact of these digital options is substantially lower than their physical alternatives. For example, Reichart & Hischier (2002) show that reading the news online is only more efficient than a printed newspaper if one considers that a large portion of the printed version is unwanted and unread. This is due to the overheads of manufacturing computers and operating the Internet, some proportion of which must be assigned to the online option. Similarly, the positive impacts of Information and Communication Technology, such as improved transport efficiency, also have negative effects through the encouragement of more travel (Hilty *et al.* 2006). How-

ever, as we continue to integrate computing into our lives the number and scope of possible applications of a generalised computing resource is likely to continue to grow. Sensor information and measurement of the infrastructure could allow us to encourage and pursue those alternatives that reduce our impact on the world. People in the developing world often live in resource-impoverished environments so a physical to digital paradigm shift has the potential to enable activities that were hitherto prohibitively expensive, and to support development whilst minimising its impact. We seek to unlock methods of wealth creation in the virtual world.

In this paper we discuss our goals of an optimal digital infrastructure and global sensing. In Section 2 we describe the concept of adaptive datacentres and show how system-level optimisations of power consumption might be achieved. In Section 3 we consider the requirements and research needs of a global sensor system. We describe our aspiration of a Personal Energy Meter (Section 4) and its possible impact on data collection and creating impetus for change. Our remaining goals considering modelling and a shift from physical to digital will be reported in a companion paper.

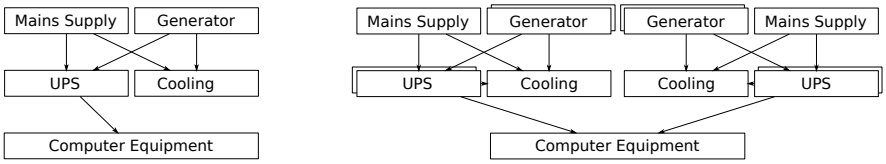
The uncertainty in predictions about future consumption levels is usually accommodated by the consideration of a number of different scenarios. The quantitative and qualitative estimates we present in this paper regarding the potential benefits of new technology consider a scenario much like the world of today. However, there is the potential for an unbounded upside to our visions. For example, sensor information and modelling the world might unlock new levels of understanding about our natural environment. This might identify actions we can take now for huge positive long-term effects. Alternatively, a massive shift of activity from physical to digital would amortise the cost of the infrastructure to an insignificant amount giving ongoing economic growth and development without incurring ecological costs. Technology has provoked massive paradigm shifts in society throughout history and it could do so again.

2. An Optimal Digital Infrastructure

The datacentre is playing an increasingly prominent role in modern computing architectures. As the size and scope of computing activities increase so do requirements for reliability and availability. Furthermore, as computing becomes mandatory our requirements for availability will increase further. In this section we consider methods for reducing overheads in high-availability datacentres and secondly for optimising the energy efficiency of our computing resources themselves.

(*a*) *Provisioning Appropriate Availability*

Many modern datacentres achieve dependability by building on high-reliability utility services. This is reflected in the traditional tiering system for datacentres (Pitt Turner *et al.*) which describes system architecture approaches for engineering availability. Figure 1a depicts a basic Tier I design which incorporates battery storage and an Uninterruptible Power Supply (UPS) to provide continued power in the event of mains failure. Increasing levels of redundancy are added to this design culminating in fault-tolerant Tier IV architectures shown in Figure 1b, which can tolerate a mechanical fault in any aspect of the system without affecting service. The



(a) Tier I (b) Tier IV

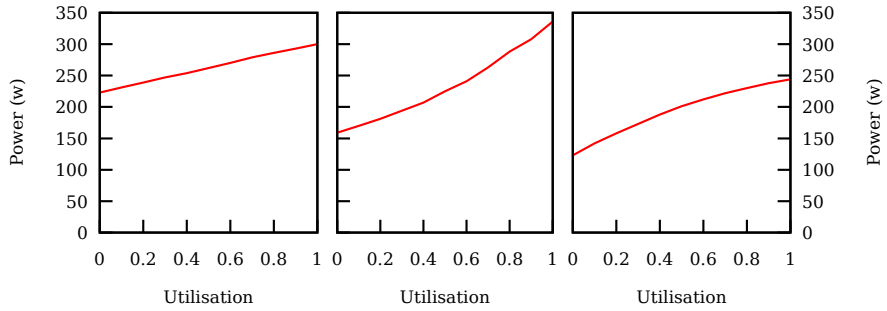
Figure 1. Simplified architecture of Tier I and Tier IV datacentres

increased availability of these architectures comes at the cost of increased redundancy. The additional support infrastructure increases the impact of construction, operation, and maintenance of the entire centre. In a classical Tier IV infrastructure every watt delivered to a server typically costs an additional watt in the support infrastructure. This gives a situation where the majority of the support infrastructure is not contributing to normal operation and exists only to cope with faults when they occur.

The approach taken in Recovery Oriented Computing (ROC) is to improve the system's recovery from failure and therefore reduce its impact (Patterson *et al.* 2002). Our goal in this respect is to decrease the restart time of services from the currently commonplace 4 hours (Pitt Turner *et al.*). This might be achieved by providing software-level support to recover and reconnect resources on startup and also by developing high-level tools to understand the interdependency between services. This means we can tolerate a reduction in redundancy, and hence an efficiency improvement, without sacrificing availability.

Fault recovery is only possible if the fault can be detected and not all faults are obvious and easy to detect. Sentient Computing (Hopper 1999) systems make use of wide-scale tracking systems, support middleware, and small battery-powered mobile devices. In this scenario faults are common but hard to locate or identify. Our technique of validation uses consistency checking for locating the source of an error in these complex interdependent systems (Rice 2007) by back-tracking through computation blocks, checking that the output values are consistent with the input values. Validation can determine when and where faults occur and often incurs only a small overhead compared with the original computation.

Datacentres are conventionally provisioned to provide a constant level of service to all applications. Instead we envision an adaptive datacentre which can exploit those situations where heterogeneity exists in the required service levels. Such a system might operate with low overheads through reduced redundancy but maintain service levels by prudently routing resources to where they are most important under fault conditions. An adaptive datacentre design would require a number of independent service systems, the sum of which are sufficient to support a fully active centre. When one of these systems fails an adaptation strategy can be determined by utilising a knowledgebase containing the resource demands of physical hardware and details of the contractual level of service required by every system. A scheduling system applies this knowledgebase to the current fault situation to determine the best allocation of remaining resources to minimize the disruption and cost of the failure. Dependency analysis between services can be exploited to determine the repercussions of shutting a component down and to cascade shutdowns to components which are rendered inoperable. This knowledgebase can also be augmented



(a) System 1 (Spec 2008a) (b) Sys. 2 (Spec 2008b) (c) System 3 (Spec 2008c)

Figure 2. Power compared to Load for real systems (SPECpower_ssj2008 results)

with expected failure rates of the components in the datacentre to estimate the overall service-level and cost of operation. One scheme for interpreting these agreements and producing an optimal assignment is to choose a target state from the space of possible options and develop an adaption plan to reach it. Administrators will require operational guarantees about the stability of the control-loop and choice of the target state is likely to require a heuristic search due to the number of possible options available.

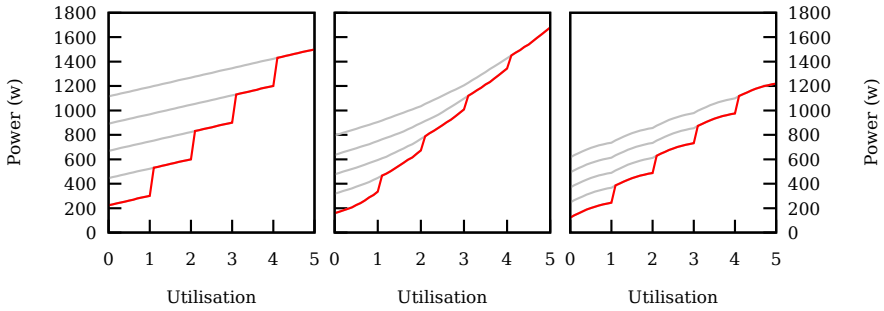
An initial example of this approach is to provide reliability guarantees at the software level. The Google File System (GFS) and MapReduce frameworks are just two systems which are designed to provide reliable operation over an unreliable infrastructure. This approach needs to be extended to adapt to capacity reductions in the power and cooling systems in the datacentre as well as failure of individual servers.

(*b*) *Energy Efficient Computing*

We now examine the power consumption of server machines themselves with the goal of energy proportional computing (Barroso & Hölzle 2007) in which server energy use scales with the amount of work done. In this section we describe how an adaptive datacentre might meet this goal.

The SPECpower benchmark tracks the relationship between power and performance under an example workload. Figure 2 shows some examples for systems running different generations of the Intel Xeon processor. Despite large variation in both maximum power and the shape of the curve on the graph we see that each of the machines shown consumes more than 50% of its maximum consumption when idle. The contribution of idle power is even more significant because server utilisation is commonly less than 50% (Barroso & Hölzle 2007). The constant presence of the machine's idle power means that this results in significant wasted energy. Indeed, Fan & Barroso (2007) demonstrate that total energy savings of at least 50% would be available given a large reduction in this idle power.

Strategies for reducing idle power can be extended to the point of switching off idle machines whilst fully utilising those which are active. Load Concentration (Pinheiro *et al.* 2003) is a technique in which incoming jobs are directed to a minimal number of busy machines allowing the idle remainder to save power. This



(a) System 1 (b) System 2 (c) System 3

Figure 3. Simulated power scaling across a cluster of systems

approach can be successful (energy consumption is potentially reduced by 30–80%) but requires specific capabilities from the server applications.

The recent resurgence in operating system virtualisation provides a platform for a pool of machines to support a number of network services. In quiet periods multiple services can run on one machine. As demand increases additional machines can be added to the pool and the services repartitioned across them. Xen is an example of virtualisation technology introduced by Barham *et al.* (2003), which partitions a single physical machine between multiple guest operating systems. A simple example of a Xen cluster architecture will contain a number of physical servers supporting some number of virtual guest operating systems running with a Storage Area Network (SAN). In a configuration such as this (where a guest's storage is always available from the SAN) Xen's live-migration system can move an active guest between two physical servers with only 250ms of external downtime and a total elapsed time of the order of 60 seconds when using a 1 gigabit per second network (Clark *et al.* 2005). It is possible with this setup to get close to the ideal of energy proportional computing when considering the energy use of the entire cluster. Figure 3 shows how power-performance curves scale for clusters of five machines and considers each of the three system types in Figure 2. The bold line shows the optimal power consumption for each utilisation level. The pale lines show variation achievable in power consumption without switching machines on or off. Moving along a pale line in the graph is rapid whereas a movement between lines requires switching machines on or off, with associated domain migration times and startup/shutdown times for the physical hardware. Compared with a single machine these clusters waste a much smaller proportion of their maximum power consumption as idle-power. The maximum power consumption of each machine remains important but as the cluster size increases the particular shape of the power scaling curve becomes less and less significant. Cole (2003) argues that power cycling a machine does not have a significant detrimental affect on its operating lifetime. Other authors note a 2% increase in failure rate of hard drives but only on drives older than 2 years (Pinheiro *et al.* 2007).

The workloads considered up to this point have been interactive workloads. They are characterised by a requirement for high availability and good response time to requests. However, there is also a significant class of non-interactive computation jobs. Particular examples include data indexing for search engines and main-stream high-performance computing tasks such as simulation batch jobs or

executing climate models. This class of jobs closely resembles the workloads of centralised, mainframe-style computing infrastructure of previous years. We can exploit the flexibility in execution of these delay-tolerant computing tasks to modulate the resource demand of our computing infrastructure. An existing example of this approach is power capping where intelligent scheduling and power-management is used to remove transient peaks in data-centre power demand (Fan & Barroso 2007).

There is a growing global impetus for the integration of large-scale renewable energy sources, wind power in particular, into the electricity grid. One of the major issues when integrating renewable energy is that of grid stability—the power output of a wind farm depends on meteorological conditions rather than conforming to a schedule that meets demand. Flexible load management of computing demand can also help in this case: unused peaks in production can be absorbed by increasing computation rates and troughs in production can be mitigated by reducing computation or shutting machines down. This is similar to dynamic demand control of consumer refrigerators which has been shown to be beneficial for grid stabiliy (Short *et al.* 2007).

Modern datacentres prioritise proximity to a source of plentiful electrical power. This reflects the economy that it is cheaper to transmit data over large distances than power. Currently, these sites target sources of reliable, stable energy supply, but an adaptive datacentre, capable of adjusting demand in response to external factors, makes this approach possible for renewable sources such as wind, tidal, and solar. Furthermore, sites for renewable energy tend to be remote, and providing high capacity power connections is difficult and expensive. Moving our datacentres to these sources of power helps alleviate this problem by using the power locally and potentially consuming energy which would otherwise be lost.

Constructing more efficient batteries and energy storage is a continuing engineering challenge. Adaptive datacentres allow computing to contribute as a "virtual battery" by selectively varying power consumption in response to the availability of surplus generation capacity. To exploit this capability fully our management systems must understand the demands of executing jobs. Computation jobs which have flexibility in their required completion times or which are amenable to interruption may well migrate around the world, tracking the generation of surplus power. Further work into the particular parameters of such a system is necessary to discover the optimum granularity for scheduling jobs and partitioning tasks. A latency map of the world would provide a useful analysis tool by showing the estimated completion time for a job depending on the targeted location for performing the computation. Ultimately, our goal is to combine the concept of computing acting as a "virtual battery" with that of turning off of all infrastructure (servers, networks, workstations and terminals) not being used for a useful end purpose.

3. The World Model

Sensor-driven applications conventionally operate with reference to a world model, which is kept up-to-date with information collected from sensors. This world model may help us not only in discovering the impacts of our activities on the world around us but also in optimising our use of energy and other resources in the existing infrastructure (Addlesee *et al.* 2001).

Careful consideration must be made as to the types of sensing and data-processing deployed in a global sensor system. One concern is that the energy consumption and environmental footprint of the sensors themselves is minimised. Sensor networks have particular processing properties that make them amenable to low-power computation: data arrives often at a regular rate and computation is inherently parallel, covering data from many sources. SpotCore is an architecture that exploits these features with a specially designed instruction set and explicit hardware support for parallelism. Eyole-Monono *et al.* (2007) demonstrate reductions in execution time and power using this platform. Careful planning of the sensor deployment and communication network will be required in order to ensure low-energy usage. The development of next-generation mobile phone networks might well be focussed on energy-efficient data collection and dissemination in addition to traditional demands for bandwidth and increased coverage.

A sensor might operate from a fixed position within the infrastructure or it might be mobile. Traditionally, static sensors have been deployed in order to exploit abundant power and communications support. An example of this occurs in the road networks of our cities where induction loop sensors are embedded in the road to detect vehicles at traffic lights. These sensors are capable of maintaining estimates of traffic queues and movements across junctions and are commonly used to improve the efficiency of the road network. More recently, improvements in technology have meant that the use of mobile sensors has become feasible. The Sentient Van project, for example, is investigating the potential for vehicle-based sensing (Davies *et al.* 2006). Sensing of traffic and travel conditions by vehicles themselves benefits from the observation that areas of high-usage for which detailed sensor data is required are naturally areas with a high density of vehicles and hence sensors. Interpretation of these sensor readings requires accurate location information (most often collected from an on-board GPS unit) and up-to-date maps of the road network. Maintaining these maps is an ever increasing burden on map makers. One solution to this problem is to collect movement traces from vehicles in the road network and post-process these traces to maintain our maps automatically. Davies *et al.* (2006) describe an algorithm both for automatic construction of a map and automatic comparison between maps to discover topology changes. Technology such as this improves the efficiency of transportation by allowing changes to the road network to be rapidly disseminated to drivers.

As a further option for sensing we consider data reports by humans acting as sensors. Utilising people as a source of sensor data has a number of benefits over conventional sensing: we are numerous and growing in number and we are self-maintaining, self-repairing and autonomous. Reports from human sensors might be qualitative: examples include local reports of illegal logging in the Brazilian rainforest, or villagers in developing worlds reporting on access to water supplies and perceived water quality. Quantitative reports are also possible from mobile sensors carried by individuals. The OpenStreetMap project (`http://www.openstreetmap.org`) is an example of this where volunteers edit and annotate GPS location traces with the aim of providing an open map resource to the world. Mechanisms for improving the reliability of reported data might include voting systems, which gain confidence from combining multiple reports, and incentive or reputation systems that increase the value (for the reporter) of a reliable report.

Data from the deployed sensor network must now be stored and processed.

Balazinska *et al.* (2007) describe their activities dealing with global scale sensor data. They identify the need for regridding data in order to cast points to a common reference frame to permit comparison and processing between datasets, and also the difficulties in representing the uncertainty in data readings and interpreting their significance. The TIME project (Bacon 2008) is constructing an open middleware for disseminating data about Cambridge's transport system. High-level descriptions of sensors are used to allow applications to interpret data without requiring tight coupling to the particular sensor instance or type.

The storage requirements for global sensing are large but quantifiable. For example, to store a single byte of sensor data for every square metre of land on the Earth requires approximately 150 terabytes. This discounts possible reductions such as compression and sparse data representation but also additional costs such as error-control coding, indexing and data about the oceans.

A centralised collection of data, particularly from human sensors, creates concerns for privacy and security. Conventionally these problems are dealt with through access control and data-dissemination policies. An alternative approach is to distribute the storage of data. This approach combined with anonymisation of data can provide an alternative for much of the complexity introduced through policy and access control. A prominent example arises from the proposed introduction of congestion charging on the UK's roads. A centralised system would inevitably collect data concerning the movements of cars within the network. However, this is not the only option. Harle & Beresford (2005) provide operating details for a system that permits peers to monitor their own compliance and so avoids the construction of a central database. The fundamental dilemmas posed by data collection technology (Gilbert *et al.* 2007) are highlighted by our considerations of energy and efficiency. A centralised database lends itself to fast and efficient processing with minimal communication costs, whereas distributed storage of data can make the system less vulnerable to privacy violation at the cost of increased overhead. Concepts such as reciprocity can make this trade-off more flexible by providing individuals with the details of people processing their data.

The demand requirements for processing sensor data will vary significantly depending on the application and incoming requests for information. It is also likely that our computing resources will experience significant variation both due to changes in adaptive datacentres and ongoing development of infrastructure. Davies *et al.* (2008) describe task-assignment in which the tasks within a computing job can be automatically assigned to a network of computing nodes. Arbitrary cost functions can be applied to this technique, producing assignments which optimise performance, power consumption, or privacy. Technology such as this allows programmers to write software in a high-level language that is automatically compiled for optimal execution on the currently available distributed resources.

An example goal of our sensing vision is the provision of a real-time data map of the world. A transportation layer on the map might show congestion on our roads and the movement and availability of public transport. An energy layer might show the state of the electricity grid indicating demand and generation levels. A water layer might show flows (and leaks) in our distribution system. Raw data sources for such a service might range from aerial imagery for observing road use or crowds in public areas to embedded sensors for recovering location information or measuring water flow and grid frequency. Collected knowledge about building

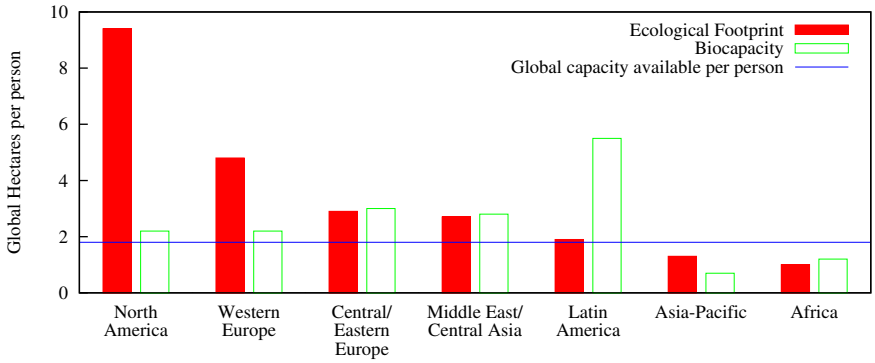


Figure 4. Per-person ecological footprints of world regions in 2002 measured in hectares of globally average biological productivity (Kitzes *et al.* 2008)

types and energy use might provide an infrared map of residential heat-loss. This example is instructive when considering the capabilities required for sensing and post-processing data. Considering a country such as the UK we need to collect consumption data from 25 million homes and 33 million registered vehicles, and monitor the use of 45 GW of electricity and 1 million litres of water per second. Data delivery must occur in a timely fashion to a centralised database whilst preserving privacy and authenticity. Some data sources such as bus locations or electricity demand require frequent updates, whilst longer timescales might be acceptable for environmental measurements such as reservoir levels. Accommodating these needs will require appropriate communications protocols and processing nodes for managing sensor data. Finally, the collected information must be presented on demand and archived for future analysis, and so support for requesting data and an interface that permits low-cost long-term archival and indexing is likely to be a further requirement.

4. Personal Energy Meters

Huge imbalances currently exist between the environmental footprint of individuals in different countries. The footprint of Western Europe is more than double its own biocapacity and over two and a half times the globally sustainable average footprint (Figure 4). Lowering the environmental footprint of our lives is a key challenge for the future of the planet. We envisage a Personal Energy Meter (PEM) which collects information about an individual's daily consumption (direct and indirect). Individualised breakdowns of the energy costs of travel, heating, water-usage and transportation of food will help us target areas for reduction in our environmental footprint. A PEM will work symbiotically with the global world model, which must feature energy as a first-class type. Data from personal activities will be uploaded to the world-model and data on the energy profile of devices and footprints of consumer goods will be downloaded to complement personal energy estimates. The data collected will not only provide useful information for analysing consumption patterns but also has the potential to help individuals identify alternatives to their current activities. For example, analysis of a commuter's PEM trace might highlight

public transport routes or potential for car sharing. Rising fuel prices might make these alternatives increasingly appealing.

The energy footprint of people in the developing world is steadily growing towards our own. However, the lack of legacy infrastructure and habits can be viewed as a potential opportunity to achieve an increased standard of living without incurring current developed-world footprints. For example, in many countries the mobile telephone network has completely supplanted any wide-scale deployment of a fixed line system, and provides communications with lower infrastructure costs and reduced environmental impact. Monitoring energy demand reported from PEMs might highlight further opportunities for technologies with similar benefits.

The design and implementation of Personal Energy Meters embodies many challenges. Effective communication with the planetary world-model must be maintained in order to provide up-to-date estimates of energy consumption. Explicit support for caching of data by publishing expected trends or validity periods for values will help to minimize this overhead. Integration of the meter itself with a mobile phone will help to minimize the energy overhead of using a PEM and also provides wide-scale communications ability. Real-time event delivery from infrastructure sensors might also exploit short range communications techniques—the PEM in your pocket might interact directly with a measurement device in the water main or the gas main as required. Finally, PEM technology might exploit the explosive growth of social networking to allow users to share and compare consumption patterns and alternatives, thus providing support for changing lifestyles and impetus for change.

5. Conclusion

Many aspects of Computing for the Future of the Planet are congruent with existing grand challenges as expressed by the National Academy of Engineering (`http://www.engineeringchallenges.org/`) and the UK Computing Research Committee (Hoare & Milner 2004). These include considerations of future energy supply, the environmental impact of modern agriculture and the improvement of urban infrastructure. The Computing for the Future of the Planet framework gives dimensions to these problems: the size and scale of a planetary world-model, though daunting, are bounded; and the factors affecting the energy consumption and impact of our activities, though many, are finite. Most importantly this framework targets research at the world in which we live outside of computing, helping to identify problems and challenges, inform their solutions and maximise their positive impact.

We have identified four specific research goals: an optimal digital infrastructure, sensing and optimising with the global world model, environmental prediction and reaction with reliable modelling, and shifting from physical to digital activities. An optimal digital infrastructure would provide us with an effective and efficient computing resource upon which we can analyse collected sensor information, model and predict future events, and potentially facilitate a shift from profligate physical activities to digital alternatives. This paper has discussed the computer science and other problems we have identified when considering system-scale power consumption for low-power computing and deploying global sensor networks.

Given the immeasurable changes undergone by computing, and caused by computing, in the last 60 years, we ask what changes might we see over the next 60

years. Computing for the Future of the Planet expresses the goal that computer technology becomes an indispensable and crucial component of our lives, society, and the environment.

We gratefully acknowledge the help of Alastair Beresford for his continued input, comments and feedback; Anthony Hylick and Sherif Akoush for their thoughts on server- and system-level power consumption; Jonathan Davies and David Cottingham for their useful feedback and corrections; Mbou Eyole-Monono for his contributions regarding low-power sensing; Rob Mullens for pointers on SPECpower; Ripduman Sohan for discussion on filesystems and other systems issues; Alan Mycroft for numerous insights and suggestions; Robert Harle and Frank Stajano for their help in the formation of our framework; and our other colleagues in the Computer Laboratory for the contributions and thoughts.

References

Addlesee, M., Curwen, R., Hodges, S., Newman, J., Steggles, P., Ward, A., Hopper, A., 2001 Implementing a Sentient Computing System. *IEEE Computer* **34**, 50–56. (DOI 10.1109/2.940013.)

Bacon, J., Beresford, A. R., Evans, D., Ingram, D., Trigoni, N., Guitton, A., Skordylis, A., 2008 TIME: An Open Platform for Capturing, Processing and Delivering Transport-Related Data, In *Consumer Communications and Networking Conference*, pp. 687–691. (DOI 10.1109/ccnc08.2007.158.)

Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T., Ho, A., Neugebauer, R., Pratt, I., Warfield, A. 2003 Xen and the art of virtualization. In *SOSP '03: Proceedings of the nineteenth ACM symposium on Operating systems principles*, pp. 167–177. (DOI 10.1145/945445.945462.)

Barroso, L. A. & Hölzle, U. 2007 The Case for Energy-Proportional Computing. *IEEE Computer* **40**, 33–37. (DOI 10.1109/MC.2007.443.)

Clark, C., Fraser, K., Hand, S., Hanson, J. G., Jul, E., Limpach, C., Pratt, I., Warfield, A. 2005 Live migration of virtual machines. *NSDI'05: Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation*, pp. 273–286.

Cole, D. 2003 *Computers and the Environment: Understanding and Managing Their Impacts*, ch. 7, pp. 131–159. Kluwer Academic Publishers.

Davies, J. J., Beresford, A. R., Hopper, A. 2006 Scalable, Distributed, Real-Time Map Generation. *IEEE Pervasive Computing* **5**, 47–54. (DOI 10.1109/MPRV.2006.83.)

Davies, J. J., Cottingham, D. N., Jones, B. D. 2006 A Sensor Platform for Sentient Transportation Research. In *EuroSSC*, LNCS 4272, pp. 226–229. (DOI 10.1007/11907503.)

Davies, J. J., Beresford, A. R., Mycroft, A. 2008 Language-Based Optimisation of Sensor-Driven Distributed Computing Applications. In *FASE 2008*, LNCS 4961, pp. 407–422.

Eyole-Monono, M., Harle, R. K., Rose, A. 2007 SpotCore: A Power-Efficient Embedded Processor for Intelligent Sensor Networks. In *Proceedings of Second International Conference on Body Area Networks*.

Hoare, T. & Milner, R. (eds) 2004 Grand challenges in computing–research. The British Computing Society. ISBN 1-902505-62-X.

Harle, R., Beresford, A., 2005 Keeping Big Brother off the road. *IEE Review* **51**, 34–37. ISSN 0953-5683.

Hilty, L. M., Arnfalk, P., Erdmann, L., Goodman, J., Lehmann, M., Wäger, A. 2006 The relevance of information and communication technologies for environmental sustainability - A prospective simulation study. *Environmental Modelling & Software* **21**, 1618–1629.

Hopper, A. 1999 The Clifford Paterson Lecture 1999: Sentient Computing. *Philosophical Transactions of the Royal Society* **358**, 2349–2358. (DOI 10.1098/rsta.2000.0652.)

Hopper, A. 2007 Computing the Planet's Future. *IET Engineering and Technology*. July 2007, p. 24

Hylick, A., Rice, A., Jones, B., Sohan, R. 2007 Hard Drive Power Consumption Uncovered. *SIGMETRICS Perform. Eval. Rev.* **35**, pp. 54–55. (DOI 10.1145/1328690.1328714)

Fan, X., Weber, W., Barroso, L. A. 2007 Power Provisioning for a Warehouse-sized Computer. *ACM International Symposium on Computer Architecture*, pp. 13–23. (DOI 10.1145/1250662.1250665.)

Gilbert, N., *et al.* 2007 Dilemmas of privacy and surveillance: challenges of technological change. The Royal Academy of Engineering.

Kitzes, J., Wackernagel, M., Loh, J., Peller, A., Goldfinger, S., Cheng, D., Tea, K. 2008 Shrink and share: humanity's present and future Ecological Footprint. *Philosophical Transactions of the Royal Society* **363**, 467–475. (DOI 10.1098/rstb.2007.2164.)

Koomey, J. G. 2006 Estimating Total Power Consumption by Servers in the U.S. and the World. Lawrence Berkeley National Laboratory, Berkeley, CA.

Miyoshi, A., Lefurgy, C., Van Hensbergen, E., Rajamony, R., Rajkumar, R. 2002 Critical power slope: understanding the runtime effects of frequency scaling. *ICS '02: Proceedings of the 16th international conference on Supercomputing*, pp. 35–4. (DOI 10.1145/514191.514200.)

Patterson, D. A., *et al.* 2002 Recovery-Oriented Computing (ROC): Motivation, Definition, Techniques, and Case Studies. UC Berkeley Computer Science Technical Report UCB//CSD-02-1175.

Pinheiro, E., Bianchini, R., Carrera, E. V., Heath, T. 2003 Dynamic cluster reconfiguration for power and performance. In *Compilers and operating systems for low power*, pp. 75–93

Pinheiro, E., Weber, W-D., Barroso, L. A. 2007 Failure trends in a large disk drive population. In *FAST '07: Proceedings of the 5th USENIX conference on File and Storage Technologies*, pp. 17–28

Pitt Turner IV, W., Seader, J. H., Brill, K. G. Tier Classifications Define Site Infrastructure Performance. White Paper. The Uptime Institute.

Reichart, I., Hischier, R. 2002 The Environmental Impact of Getting the News: A Comparison of On-Line, Television, and Newspaper Information Delivery. *Journal of Industrial Ecology* **6**, 185–200

Rice, A. 2007 Dependable systems for Sentient Computing, ch. 5. Ph.D. Thesis, University of Cambridge (UCAM-CL-TR-686).

Roselli, D. S., Lorch, J. R., Anderson, T. E. 2000 A Comparison of File System Workloads. *USENIX Annual Technical Conference*, pp. 41–54.

Short, J. A., Infield, D. G., Freris, L. L. 2007 Stabilization of Grid Frequency Through Dynamic Demand Control. *IEEE Transactions on Power Systems* **22**, 1284–1293. (DOI 10.1109/TPWRS.2007.901489.)

Spec 2008a SPECpower_ssj2008 results for Supermicro 6025B-TR+. Standard Performance Evaluation Corporation. `http://www.spec.org/power_ssj2008/results/res2008q1/power_ssj2008-20080115-00029.html`.

Spec 2008b SPECpower_ssj2008 results for Intel Platform SE7520AF2 Server Board. Standard Performance Evaluation Corporation. `http://www.spec.org/power_ssj2008/results/res2007q4/power_ssj2008-20071129-00015.html`.

Spec 2008c SPECpower_ssj2008 results for Proliant DL180 G5. Standard Performance Evaluation Corporation. `http://www.spec.org/power_ssj2008/results/res2008q1/power_ssj2008-20080116-00033.html`.

Williams, E. 2004 Energy Intensity of Computer Manufacturing: Hybrid Assessment Combining Process and Economic Input–Output Methods. *Environmental Science and Technology* **38**, 6166–6174.

In Phil. Trans. R. Soc. A 366(1881):3685–3697, 2008
Please see http://journals.royalsociety.org/content/c5241623l1g077nn/