CHAPTER 6

Concluding Remarks

Theo Lynn and John P. Morrison

Abstract Traditionally, access to high performance computing was restricted by architectural complexity, availability of trained personnel, and budgetary issues. At the same time, research suggests that existing measures for greater data centre energy efficiencies will reach theoretical and practical limits in the near future. This concluding chapter briefly discusses the potential of (i) cloud computing to disrupt the high performance computing sector, and (ii) new heterogeneous cloud architectures, based on the concepts of self-organisation, self-management, and the separation of concerns, to disrupt extant cloud resource management approaches.

Keywords Disruptive innovation • Cloud computing • High performance computing • HPC in the cloud

T. Lynn (✉)
Irish Centre for Cloud Computing (IC4), Dublin City University, Dublin, Ireland
e-mail: theo.lynn@dcu.ie

J. P. Morrison
Department of Computer Science, University College Cork, Cork, Ireland
e-mail: j.morrison@cs.ucc.ie

T. Lynn et al. (eds.), *Heterogeneity, High Performance Computing, Self-Organization and the Cloud*, Palgrave Studies in Digital Business & Enabling Technologies,
https://doi.org/10.1007/978-3-319-76038-4_6

Clayton Christensen, in his seminal study on the disk drive industry, identified two types of technological change. Sustaining technologies sustained the industry's rate of improvement in product performance and ranged in difficulty from incremental to radical, whereas so-called disruptive innovations redefined performance trajectories and consistently resulted in the failure of the industry's leading firms (Christensen 1997). Cloud computing continues to transform, and democratise access to, the use of information and communications technology infrastructure. Organisations of all sizes and sectors, as well as the general public, are able to exploit the advantages of the agility and scalability (up and down) inherent in cloud computing to work more efficiently, reduce Information Technology (IT) costs (including IT capital expenditure, maintenance and support costs, and related environmental costs), support resilience and business continuity, and growth (Hogan et al. 2011; Low et al. 2011; Buyya et al. 2009; Leimbach et al. 2014). This book is about disruptive potential—the (i) the potential of cloud computing to disrupt the high performance computing (HPC) sector and (ii) the potential of a new heterogeneous cloud architecture based on the concepts of self-organisation, self-management, and the separation of concerns to disrupt extant cloud resource management approaches.

For a significant portion of the last half-century, HPC exploited relatively established trajectories of performance; single-thread processor clock frequency was viewed as the main driving factor behind increasing computational performance. Manufacturers of such processors, and Intel in particular, delivered consistent improvements in performance until hitting a scientific "power wall" for single-core processors at the turn of the century. With the levelling off of single-thread processor performance, the industry sought to sustain performance trajectories by combining multiple Central Processing Unit (CPU) cores on one chip to achieve performance gains. While multi-core architectures achieve performance gains, efficient parallel computation on multiple cores provided discrete challenges for the HPC end user community. More recently, the use of different types of processor has been exploited to address this issue. As different compute resources can have different properties, applications with diverse characteristics can be executed quicker and more efficiently using these processors.

Heterogeneous architectures support these specialist processors as co-processors to a host processor; the host processor can complete one instruction stream, while the co-processor can complete a different

instruction stream or different type of stream (Eijkhout et al. 2016). While such heterogeneous resources can provide new measures of performance, for example, energy efficiency, both technically and culturally the HPC community remains focused on maximising the (effective) processing speed of a given architecture to orders of magnitude greater than general-purpose computing. Whereas each evolution of the processor architecture was relatively novel in the context of difficult HPC applications, it was not disruptive. To paraphrase Christensen (1997), the customers of the leading HPC supplier led them towards these achievements. These sustaining technologies did not precipitate failure by incumbents or significant changes in the HPC industry structure. Size still matters. The HPC community remains dominated by a relatively small number of suppliers catering for a relatively small number of large organisations requiring significant investments in infrastructure. For the most part, access to HPC remains restricted by architectural complexity, availability of trained personnel, and budgetary issues (Intersect360 Research 2014).

In the last few years, cloud service providers (CSPs) have sought to enter the HPC market; however, HPC has remained one of the smallest segments in the market. This can be explained by both technical and cultural perceptions on the nature of HPC and the efficacy of cloud computing architectures to deliver high performance. From a technical perspective, many HPC workloads are not ready to run on today's cloud architectures, and provisioning of HPC clusters in the cloud still typically requires deep IT knowledge. Similarly, many in the HPC community do not believe a general-purpose distributed architecture designed for multi-tenancy, horizontal scaling, and minimal interference with physical infrastructure can deliver the performance expectations for HPC. And this may be correct.

However, there are classes of HPC users who do not need maximum performance, and this goes to the core of the disruptive potential of cloud computing for HPC. Cloud computing creates new markets and value networks for organisations (and individuals) who cannot afford or cannot gain convenient access to traditional HPC infrastructure such as supercomputers, who have loosely coupled workloads that can be scaled horizontally, and/or have pent-up HPC demand and find it difficult to burst capacity for overflow or surge workloads with their existing HPC infrastructure. Given the impact HPC has on scientific discovery and innovation, dramatically increasing access and use of HPC through the cloud to this wider community of low-end consumers or non-consumers has the potential to drive significant societal and economic impact.

At the same time, it is questionable whether the economic model of conventional hyperscale cloud computing is sustainable in the long term. Not from a business or technology perspective but from an environmental perspective. The IT sector accounts for a significant portion of global electricity with some estimates at approximately 7% (Corcoran and Andrae 2013). Data centres have an extremely energy-intensive profile. For example, a study conducted for the US Department of Energy estimates that data centres consume 10–50 times the energy per floor space of a typical commercial office building and collectively (Darrow and Hedman 2009). In 2014, data centres accounted for 1.8% of total US electricity consumption driven by increased Internet usage and the rise of cloud computing (Shehabi et al. 2016). Research suggests that the data centre sector, and hyperscale data operators specifically, has taken significant measures to improve energy efficiency including increasing server productivity and utilisation and efficiency improvements in storage, network, and data centre infrastructure operations such as cooling (Shehabi et al. 2016). Despite these initiatives, the environmental impact of Information and Communications Technologies (ICT) operations, data centres, and cloud energy usage remains a significant concern and increased focus of policy makers and civic society.

Research suggests that existing measures for greater data centre energy efficiencies will reach theoretical and practical limits in the near future, and therefore cloud computing especially needs to look beyond its current model of using one-size-fits-all hardware towards optimising hardware for specific workloads (Shehabi et al. 2016). Such optimisation is central to the heterogeneous cloud; however, such a vision for cloud computing increases the complexity of managing cloud infrastructure dramatically. As such, a new paradigm for cloud computing architectural design is required.

This book presents one possible architectural design, CloudLightning, for managing heterogeneous clouds based on self-organisation, self-management, and the separation of concerns. CloudLightning is a fundamentally different architecture to the homogeneous cloud platforms prevalent today. Specifically, it both accommodates workload variation through optimised heterogeneous hardware and hides this complexity from enterprise application developers and end users, thus providing a different package of attributes including not only hardware performance but energy efficiency, ease of management, and ease of use as well. CloudLightning's disruptive potential is the new performance trajectory that such attributes create.

References

Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems, 25*(6), 599–616.

Christensen, C. M. (1997). *The innovator's dilemma: When new technologies cause great firms to fail.* Cambridge, MA: Harvard Business School Press.

Corcoran, P., & Andrae, A. (2013). *Emerging trends in electricity consumption for consumer ICT.* Tech. Rep., National University of Ireland, Galway, Connacht, Ireland. Retrieved October 24, 2016, from https://aran.library.nuigalway.ie/xmlui/handle/10379/3563

Darrow, K., & Hedman, B. (2009). *Opportunities for combined heat and power in data centres.* Arlington, VA: ICF International.

Eijkhout, V., van de Geijn, R., & Chow, E. (2016). Introduction to high performance scientific computing. *Zenodo.* https://doi.org/10.5281/zenodo.49897

Hogan, M., Liu, F., Sokol, A., & Tong, J. (2011). *NIST cloud computing standards roadmap.* NIST Special Publication, 35.

Intersect360 Research. (2014). *Worldwide high performance computing 2013: Total Market Model and 2014–18 forecast.* Sunnyvale, CA.

Leimbach, T., Hallinan, D., Bachlechner, D., Weber, A., Jaglo, M., Hennen, L., et al. (2014). Potential and impacts of cloud computing services and social network websites (STOA Cloud Computing—Study). Retrieved October 26, 2017, from http://www.europarl.europa.eu/RegData/etudes/etudes/join/2014/513546/IPOL-JOIN_ET(2014)513546_EN.pdf

Low, C., Chen, Y., & Wu, M. (2011). Understanding the determinants of cloud computing adoption. *Industrial Management & Data Systems, 111*(7), 1006–1023.

Shehabi, A., Smith, S. J., Horner, N., Azevedo, I., Brown, R., Koomey, J., et al. (2016). *United States data centre energy usage report LBNL-1005775.* Berkeley, CA: Lawrence Berkeley National Laboratory.