



Using FPGAs to Minimise Data Centre Power Consumption

Telesoft White Papers

Christian Feest
22nd September 2015



"In 2013, U.S. data centres consumed an estimated 91 billion kilowatt-hours of electricity - enough electricity to power all the households in New York City twice over - and are on-track to reach 140 billion kilowatt-hours by 2020." - National Resources Defence Council Report^[4]

Introduction

The rising popularity of cloud computing, virtualisation and remote storage is fuelling increased investment in data centres. By 2018 global data centre traffic is expected to reach 8.6 zettabytes - the equivalent of 9 trillion hours of high definition video streaming - which is more than double the 2013 total of 3.1 zettabytes^[2].

As data centre traffic rates have increased, interest in minimising the costs, both financial and environmental, has gathered momentum. It would take the annual output of 50 large coal-fired power stations to provide the projected 140 billion kilowatt-hours of electricity required to power U.S. data centres in 2020 and the environmental impact of this has been compared with that of air travel. These costs not only present a challenge for a sustainable future, they also represent a significant financial burden on data centre operators; a burden compounded further by rising energy prices.

With the spotlight on these issues methods of reducing energy expenditure can be divided, broadly, into two categories: more efficient servers and more efficient systems for cooling them. This white paper focuses on the former method by looking at how a large proportion of data centre processing burden can be off-loaded to hardware accelerator cards to reduce host CPU load and cut overall energy requirements.


Global Data Centre Traffic Estimation

2018
8.6 zettabytes
Global data traffic

Cloud data centre traffic will quadruple between 2013 and 2018

By 2020 US Data Centres

2020
140
billion kilowatt-hours
Electricity required



It would take...
50
Coal Fired Power Plants



This will significantly increase costs and environmental impact



The Challenge

For most operators the operating expenditure (OPEX) associated with running a data centre is greater than the capital expenditure (CAPEX). For many of these operators the electricity costs alone over the lifetime of a piece of equipment exceed the cost of the equipment itself. Improving data centre efficiency thus has significant advantages for operators who can enjoy potentially huge reductions in OPEX whilst helping the environment and acquiring desirable green credentials that can be publicised to maintain a positive brand image.

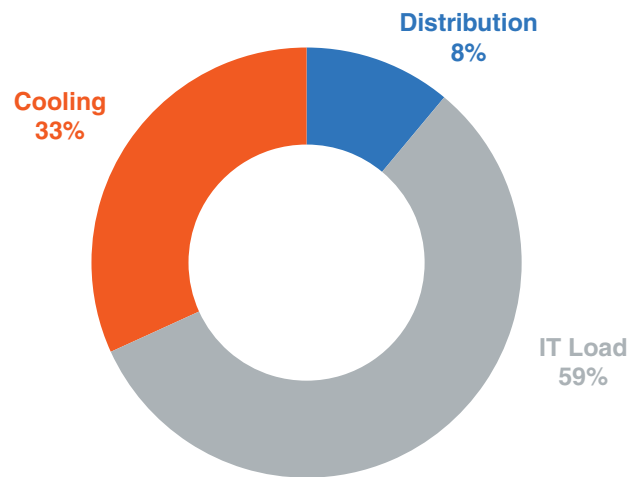


Figure 1: Average Data Centre Power Consumption by Component^[3]

The most commonly used metric for measuring data centre efficiency, Power Usage Effectiveness (PUE), expresses IT equipment power usage as a proportion of total power usage. An ideal PUE of 1.0 would mean every watt of electricity used in data centre operation is used to power servers with no energy wasted on cooling, lighting, etc. Despite its ubiquity, this metric only tells part of the story. A water cooled data centre in Antarctica that uses the most advanced environmental control technology could achieve a PUE of 1.1, for example, meaning that for every 10 watts of energy used in computation only 1 watt of energy is needed for non-computational purposes. However, if the IT equipment being cooled is itself massively inefficient and mostly underutilised, the data centre would still be wasting huge amounts of energy unnecessarily.

A properly efficient data centre would thus not only have a high PUE but also a high Performance per Watt (PPW). PPW measures the amount of computation that can be performed for each watt of power consumed and with IT resources accounting for nearly 60% of average data centre power usage this represents the largest potential area to reduce energy consumption.

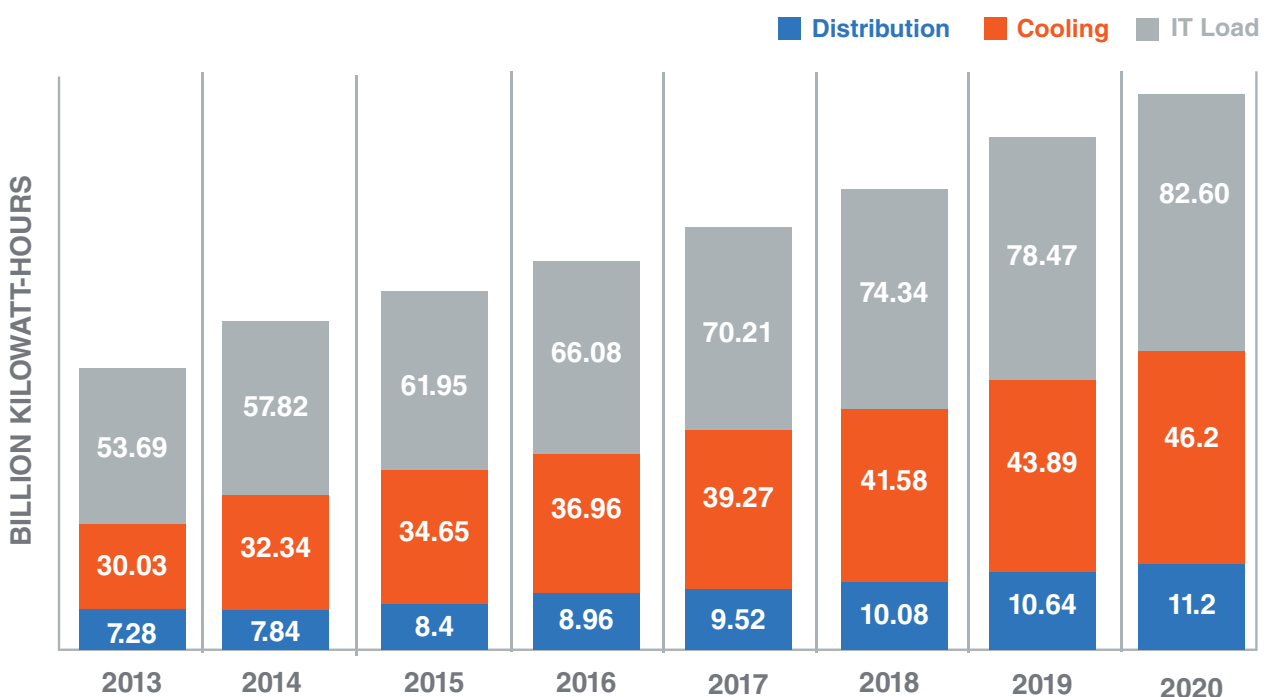


Figure 2: Projected data centre power consumption in the U.S.

Hardware Acceleration

“It’s going to be absolutely critical that future data centre designs include GPUs or FPGAs [...] You can get at least an order-of-magnitude improvement.” – Jason Mars, Assistant Professor of Computer Science, University of Michigan^[6]

With the average data centre server operating at just 12-18% of capacity^[4], the easiest way to improve overall PPW for many operators is to identify and decommission unused servers and power off unnecessary equipment during quiet times. However, for operators who have already taken steps to minimise the number of comatose servers, there is still huge scope to dramatically reduce energy requirements by improving the efficiency of current equipment. Further, as the range and complexity of services that data centres are expected to provide increases, inefficient hardware and the resultant high OPEX will drain resources, hindering the introduction of new and improved services that keep operators ahead of their competitors.

As the end of Moore’s law approaches and server performance improvement rates slow down, operators are looking at alternative computing architectures to enable continued and consistent improvements to services that remain economically viable. With power draw being the limiting factor for performance gains in traditional servers, Microsoft’s pioneering Catapult research program is leveraging Field-Programmable Gate Array (FPGA) technology to shoulder a large proportion of the processing burden whilst reducing power consumption. FPGAs, hardware devices that can be programmed to carry out bespoke processing tasks, offer significantly greater PPW compared to traditional CPUs for many applications. The Catapult project, initially a pilot system comprising 1632 servers, was designed as a system to speed up Microsoft’s Bing web ranking algorithm. The ability of FPGAs to process certain Bing algorithms 40 times faster than the equivalent CPU meant Bing’s web ranking could be achieved using approximately half the number of servers that were required previously and the success of the Catapult architecture has led to its implementation in live data centres.

“The Catapult architecture is really much more general-purpose, and the kinds of workloads that [...] can be dramatically accelerated by this are much more wide-ranging.” – Peter Lee, Corporate Vice President, Microsoft Research^[5]

The potential range of applications for hardware acceleration using FPGAs is not just limited to web ranking either. Digital assistants, such as Apple’s Siri, Google Now and Microsoft Cortana, require voice recognition and natural language processing across multiple servers. The hundreds of small but parallel calculations required by digital assistants makes them perfectly suited to implementation in FPGA. As digital assistants become increasingly popular and the complexity of tasks they can perform increases it will be increasingly untenable to use ordinary servers to accommodate them. Offloading the computationally intensive aspects of these and similarly parallel applications to dedicated hardware acceleration platforms will allow operators to increase the processing power of their data centres in an economically and environmentally efficient way.

Examples of tasks that are similarly suited to hardware acceleration using FPGAs include:

- Image recognition and classification
- Encryption and decryption
- Video applications (e.g. encode and decode)
- Cloud security
- Load balancing
- Internet key exchange
- Deep learning and neural networks

MPAC-IP 7000 Series

The MPAC-IP 7000 Series; can accelerate processing of a range of applications from intrusion detection and prevention (e.g. Snort, Bro, Suricata) and high frequency trading platforms to traffic discrimination for differentiated services. The MPAC-IP series is available in 4x10Gbps and 2x100Gbps to reduce host CPU power requirements and accelerate processing in a single high density, low footprint solution. Connected to COTS servers via PCIE 3.0 and controlled using industry standard APIs, integration has never been simpler. To find out more, call us today, or visit our website www.telesoft-technologies.com

Conclusion

The rising popularity of cloud computing and the increased demands this puts on data centres presents numerous challenges. Not only are more and more features supported by the cloud but the computational complexity of these features is also increasing, meaning more power is required to provide them. As cloud computing continues to gather momentum, operators of data centres can no longer rely on Moore's law to provide the consistent improvements in computational efficiency required to keep up with market developments in an economically and environmentally cost-effective way.

Hardware acceleration, through the use of FPGAs, provides an alternative computational architecture that is well-suited to projected data centre demands. Many data centre operations are parallelisable and thus can be accelerated through the use of FPGAs to increase PPW. As Moore's law no longer holds true for traditional CPUs the next generation of data centres can offload large proportions of their processing burden to FPGAs, as Microsoft have demonstrated in the Catapult research program, to benefit from the increased performance and reduced operating costs that hardware acceleration can bring.

Sources

1. A. Putnam, A.M Caulfield, E.S Chung, D. Chiou, K. Constantinides, J. Demme, H. Esmailzadeh, J. Fowers, G.P Gopal, J. Gray, M. Haselman, S. Hauck, S. Heil, A. Hormati, J. Kim, S. Lanka, J. Larus, E. Peterson, S. Pope, A. Smith, J. Thong, P.Y Xiao and D. Burger: A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services, Microsoft. Retrieved 08/09/2015 from <ftp://ftp.cs.utexas.edu/pub/dburger/papers/ISCA14-Catapult.pdf>
2. Cisco Global Cloud Index: Forecast and Methodology, 2013-2018, Cisco. Retrieved 08/09/2015 from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.pdf
3. Power Management in the Cisco Unified Computing System: An Integrated Approach, Cisco. Retrieved 08/09/2015 from http://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/unified-computing/white_paper_c11-627731.pdf
4. J. Whitney and P. Delforge: Data Centre Efficiency Assessment, National Resources Defense Council. Retrieved 08/09/2015 from <http://www.nrdc.org/energy/files/data-center-efficiency-assessment-IP.pdf>
5. R. McMillan: Microsoft Supercharges Bing Search with Programmable Chips, Wired 16/06/2014. Retrieved 08/09/2015 from <http://www.wired.com/2014/06/microsoft-fpga/>
6. C. Metz: Microsoft Knows Exactly Where Intel's Future Is, Wired 05/06/2015. Retrieved 08/09/2015 from <http://www.wired.com/2015/06/microsoft-knows-exactly-intels-future/>



› **Headquarters**

Telesoft Technologies Ltd,
Observatory House, Stour Park
Blandford DT11 9LQ UK
t. +44 (0)1258 480880
f. +44 (0)1258 486598

[e. sales@telesoft-technologies.com](mailto:e.sales@telesoft-technologies.com)

› **Americas**

Telesoft Technologies Inc
125 Townpark Drive, Suite 300
Kennesaw, Georgia, GA 30144 USA
t. +1 770 454 6001

[e. salesusa@telesoft-technologies.com](mailto:e.salesusa@telesoft-technologies.com)

› **India**

Telesoft Technologies Ltd (Branch Office)
Building FC-24 Sector-16A, Noida 201301
Uttar Pradesh, INDIA
t. +91 120 466 0300
f. +91 120 466 0301

[e. salesindia@telesoft-technologies.com](mailto:e.salesindia@telesoft-technologies.com)

www.telesoft-technologies.com