Holistic Data Centres: Next Generation Data and Thermal Energy Infrastructures

Paul Townend, Jie Xu
School of Computing
University of Leeds
Leeds, UK
{p.m.townend, j.xu} @ leeds.ac.uk

Jon Summers, Daniel Ruprecht, Harvey Thompson
Dept. of Mechanical Engineering
University of Leeds
Leeds, UK
{j.l.summers, d.ruprecht, h.m.thompson} @ leeds.ac.uk

Abstract— Digital infrastructure is becoming more distributed and requiring more power for operation. At the same time , many countries are working to de-carbonise their energy, which will require electrical generation of heat for populated areas. What if this heat generation was combined with digital processing?

Keywords—holistic data centers; exergy; characterization; heat

I. Introduction

Data centers are “21st century factories” that process electrical power into digital services and generate enormous quantities of waste heat that requires similar power to remove. The greenhouse gas (GHG) emissions from the world data centre sector (80MtCO2e as of 2007) are set to quadruple by 2020 [1], and the increased growth and energy use within data centers are a direct threat to the sustainability of the digital economy and achieving the agreed GHG emission targets of 20% of 1990 levels by 2050 (Kyoto Protocol).

Whilst it is clear that the digital infrastructure brings reductions in GHG emission levels, there has been little attention to the greening of the data centers that underpin such savings. Currently the average utilization of a data centre is as low as 10% [2]. Suppliers to the data centre industry are starting to address aspects of energy efficiency, e.g. layout for efficient cooling [3], efficient IT components and in-workload scheduling [4], consolidation [5], resource throttling [6]. However, the data centre is a mission-critical facility with implicit requirements on availability, security, capacity and robustness, which yield design and operational constraints. The data centre industry is conservative in adopting energy saving measures without understanding the implications of the technological choices on their business practices.

Although there has been a rapid growth in digital infrastructure, the data centre industry is still young, fragmented and immature, lacking in scientifically-rigorous, independently peer-reviewed tools and methodologies for whole system design, construction, operation and maintenance. A major contribution to these problems are the inherent multidisciplinary nature and diversity of solutions, which requires in-depth knowledge and understanding of the technological (software stacks and ICT hardware) as well as the environmental (mechanical, electrical and building services engineering) systems to develop a holistic overview and system-level understanding of a data centre. The role of business decisions is also significant (for example in defining the uptime of the facility; identifying energy efficiency as an opportunity rather than a risk) and must be considered within a whole-systems approach to energy-efficient data centers.

We introduce a high-level system model - shown in figure 1 - that identifies the core physical, virtual, power, and business process components of a data centre, and the key inter-connections between them. From this starting point, we aim to:

1. Characterize data centers in terms of the IT, physical and the behavioural aspects of energy-efficient operation. This will be achieved through k-means clustering analysis (to automatically categorize users, workloads, etc.) and distribution fitting (giving an approximation of the behaviour of each identified category with respect to two arbitrary parameters).

2. Establish a system-level methodology that views a data centre as a complex System of Systems (SoS) for whole system design, construction, operation and maintenance. This enables a holistic view of data centers that elucidates the effects of location, risk, reliability, redundancy and resilience on end-use energy demand and whole system lifecycle costs.

3. Explore and develop novel new methods to both mitigate and harness the heat caused by computation within in a data centre – notably develop a framework for exergy efficiency.

For this poster paper, the focus is on the last of these aims; we introduce the concept of Exergy efficiency, and discuss some of the software technologies that can be used to drive it.

II. Exergy Efficiency

Currently, ICT facilities deal with waste heat by a combination of cooling (through a variety of different hardware techniques such as air and liquid cooling) and rejecting thermal energy into the environment; both of these methods result in further significant economic and environmental cost. Exergy efficiency is a radical new approach to managing waste heat; it proposes that the efficiency of ICT devices can be greatly improved if the generated (and previously wasted) heat can instead fulfil some useful purpose (for example, small clusters located near a swimming pool could be used for processing; the heat generated could then be used to heat the pool). In essence, the locally consumed data will become the fuel for thermal

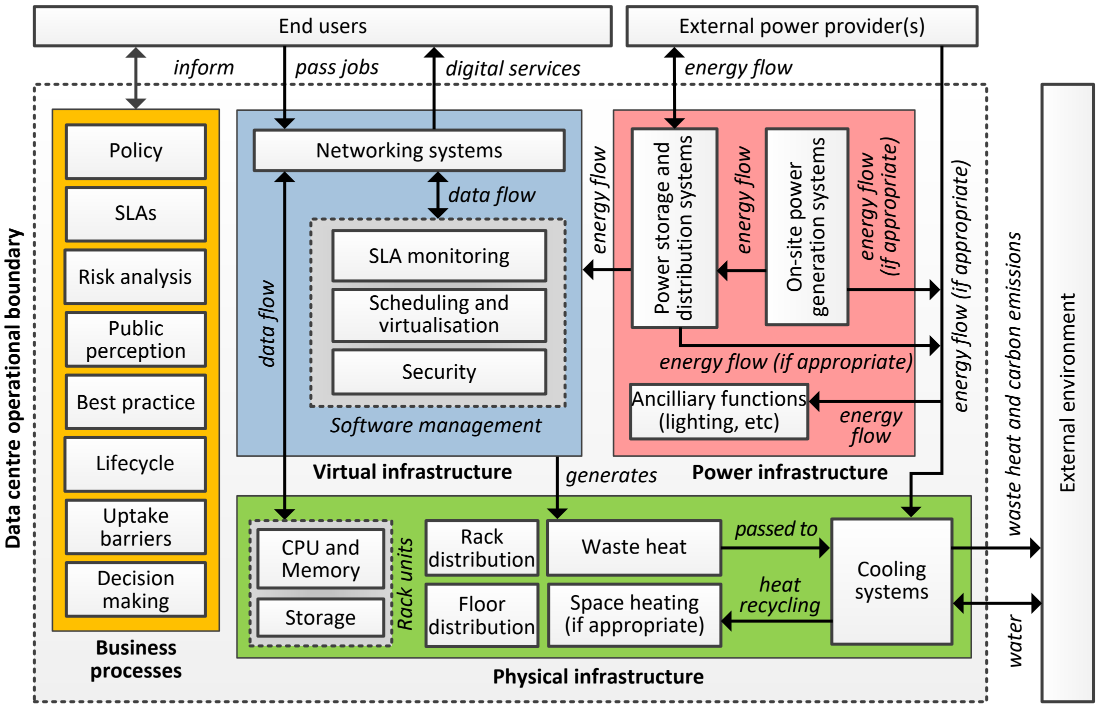

Figure 1: High-level Data Centre System Model

energy, whether it be heating or cooling. To achieve this aim, there needs to be a scheduling infrastructure in place that dynamically reschedules jobs to areas that require heat or power.

Accurate resource management will necessitate accurate workload characterizations, derived through the process described in section 1. Upon knowing the predicted characteristics of an incoming workload, a scheduler can determine the approximate thermal energy that will be created by processing the workload, and intelligently route the workload to an appropriate local cluster. To enhance scalability, each local cluster can utilize its own scheduler for further load balancing

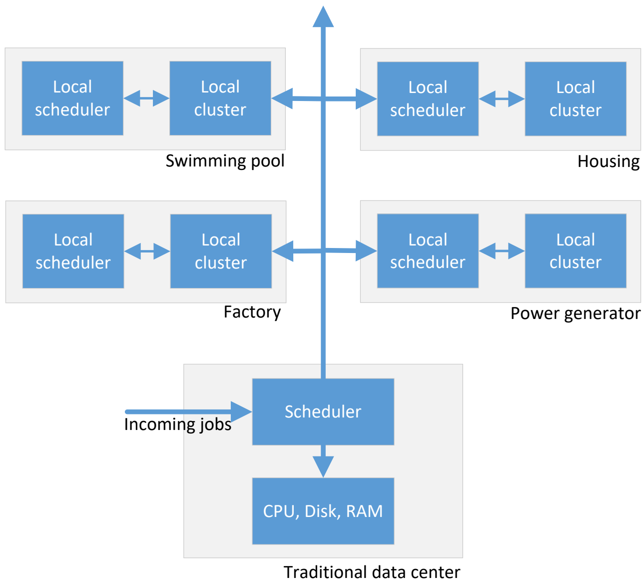

Figure 2: Example applications of Exergy Efficiency

and performance/failure management. At all times, characterizations of individual node characteristics (performance, load, hardware architecture, etc.) will need to be considered in order to ensure necessary Service Level Agreements (SLAs) are maintained.

Such an approach has enormous potential impact worldwide, but requires significant multi-disciplinary effort to achieve, combining heat, cooling, and computation. Two current trends facilitate this shift in paradigm. First, the trend away from massive centralized data centers toward distributed but ubiquitous smaller computing devices and second the trend for new materials, like Silicon Carbide or Gallium Nitride that could offer greater speed but operate at higher temperatures.

By investigating the concept of ``data as fuel'', we aim to explore a potentially disruptive approach that could lead to an enormous increase in energy efficiency. Supporting the trend toward more distributed computing resources will furthermore help to make this infrastructure more resilient.

REFERENCES

[1] Energy Efficiency: A compelling Global Resource, www.mckinsey.com 2010.

[2] A. Greenberg, J. Hamilton, et al., “The Cost of a Cloud: Research Problems in Data Center Networks” in Computer Communication Review, 39(1): p. 68-73. 2009.

[3] M. Martin, et al, “High-density heat containment” in ASHRAE Journal pp38-43, 2007.

[4] S. Albers, “Energy-efficient Algorithms”, in Communications of the ACM, 53(5), 86-96, 2010.

[5] S. Srikantaiah et al, “Energy aware consolidation for cloud computing”, in Cluster Computing, 12, pp. 1–15, 2009

[6] M. Stillwell et al, “Resource allocation using virtual clusters”, in Proc. 9th IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2009), Shanghai, China, pp. 260–267, 20