Towards a Holistic Data Center Simulator

Sriram Sankar, Aman Kansal, Jie Liu
Microsoft Corp.
One Microsoft Way, Redmond, WA 98052
{sriram.sankar, kansal, liuj}@microsoft.com

Data center (DC) design has become increasingly important with the rapid growth of cloud computing and online services. The rapid growth rate makes them a significant consumer on the energy grid. Differences in environmental operating conditions, energy price and availability, network bandwidth and latency, as well as unpredictable user demand pose significant challenges for determining the right size, density, and energy sources for data centers. Data from real data centers is often proprietary and severely limits academia and research institutions from addressing these challenges. Building a data center testbed for research is not only cost prohibitive (e.g., a 1 MW datacenter costs approximately $10 Million- $22 Million [1]) but is also difficult to continually upgrade or explore diversified technologies and industry practices.

Existing modeling, design methodologies and tools are not capable of capturing the scale and heterogeneity in complex systems like data centers. To effectively model performance, energy consumption, energy technologies, network, server trends, failure recovery, and varied operational scenarios, we propose coordinated research efforts to build a DC level full system modeling and simulation platform that enables researchers to investigate multiple DC design aspects for energy and resource efficiency.

Design challenges

Designing an effective DC simulator poses interesting research challenges, including the following:

1. Multiple Interacting Systems: Multiple physical and computing systems interact in a DC. For instance, Microsoft's Chicago datacenter can hold 300,000 servers, 11 diesel generators each supplying 2.8 MW of power, 11 electrical substations and power rooms, 12 chillers each with a capacity of 1260 tons and arrays of network switches and cables [8]. This inter-dependent infrastructure must be managed correctly to operate the DC at optimum levels.

2. Multiple Time Scales from Component to Systems: The different interacting systems have different time constants. For instance, a utility failure prorogates at the speed of light, disk access times in backend servers must be modeled at millisecond scales, user demand variations at several seconds to minutes, whereas an adiabatic cooling system effect only manifests itself after several hours (e.g, a cooling failure at night might not impact operation till servers heat up during the day). Even longer scales are required for solid state drives where workload induced endurance issues take months or years to manifest.

3. Multiple System Granularities: It is important to understand the system granularity needed to simulate DC scenarios. For example, to understand workload dynamics, does capturing CPU utilization suffice, or do we need to simulate workload behavior at the L2, L3 level caches? Do we simulate power behavior of each individual component in a motherboard, or do we simulate system level power based on utilization? Do we need full air dynamics simulator to understand cooling efficiency? There are several such granularity questions that need to be answered.

4. Design time vs Run-time simulations: Certain scenarios need design time answers while others are only relevant during operation. For instance, the total cost building and operating a datacenter is impacted more by the power provisioning time decisions than run-time power consumption cost [1]. On the other hand, during run-time, we need to consider performance impact to application running at peak during the time power is being capped.

5. Statistical Summary vs Patterns: Typically, the most stringent design requirements come from fault tolerance due to high reliability requirements. Corner cases and specific sequences of events are important and just summary statistics might not suffice.

6. Metrics and Scale: Several metrics are used to quantify DC operations including Availability, Utilization, Capacity, Performance, Cost and Power. Even for performance, there are many metrics like Queries per second, Throughput, IO per second, Transactions per second, FLOPS, MIPS, Mbps, etc.. The simulation platform must be able to handle these metrics at scale. For example, in our experiments of collecting performance counters from production data centers, the system can easily generate 1TB data per day.

Preliminary Results

While the challenges are far from addressed, we present preliminary work towards some aspects of a DC simulator. A fundamental capability to run a simulation is to incorporate desired workload in the analysis. It is impractical to record every aspect of a workload such as fine grained resource usage on CPU, memory, network, storage, power and so on for multiple interesting workloads. More importantly, such fine grained recordings limit flexibility because they can only allow a specific workload run to be recreated and do not allow simulating design changes such as introducing a faster network or different storage technology. Hence, the workload should be abstracted out in a manner that preserves its requirements but allows exploring the performance of different design options. We may create such abstractions for different data center components that a workload uses. We show several examples that we have created.

Storage: An application's functionality, and consequently the data needs are preserved across multiple data center designs, and hence an appropriate abstraction of storage use is to model the specific data elements accessed by an application. We capture statistical properties of the workload at different selectable spatial and temporal scales required for different design scenarios [2]. For example, coarse granularity can simulate large server clusters, while finer grained incarnations allow exploring different RAID array configurations and solid state drives.

Network: Communication patterns between different distributed software components of an application can greatly affect the design of the network capacity required and resultant performance. To capture an application's requirements, we model the network traffic in both spatial and temporal dimensions. We again use a statistical hierarchical state-based model (Figure 1) that allows re-creating such traffic at different scales. Figure 2 shows spatial patterns of inter-server network traffic for Websearch, across a server cluster, and its simulated recreation, and shows that the model follows the original [3].

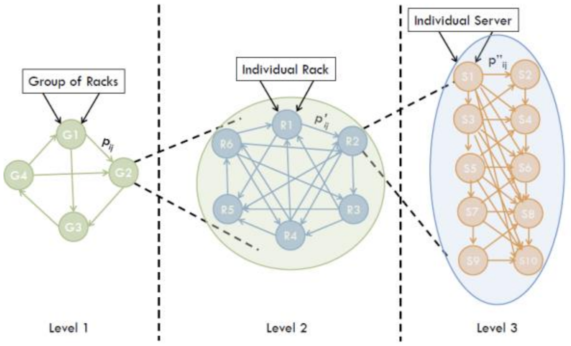

Figure 1: State based model with multiple levels

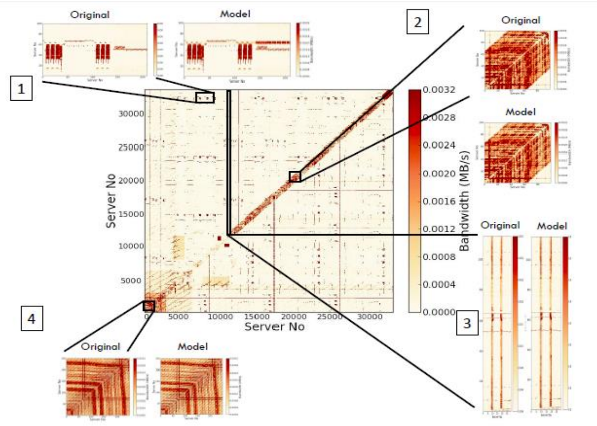

Figure 2: Network model comparison with original

Prior research has addressed other specific challenges that will aid the design of the proposed simulator. GDCSim [4] models energy and cooling mechanisms. BigHouse [7] uses stochastic queuing to model power and load balancing studies. MDCSim [6] simulates the application layer with queuing models. Server queuing models were used in [5]. Most prior works targeted specific systems or were based on queuing models which did not provide a detailed simulation of sub-system characteristics. Our proposal is to combine sub-system level simulation, while preserving inter component timing and interaction details through a hierarchical state based model that allows for tradeoff between detail and simulation time.

Energy Modeling. The key connection between computing and network activities and physical power and cooling activities is energy. We have developed a machine learning based approach to map performance counter values to hardware *and software* energy consumption and have successfully applied in server power capping.

Ultimate Goals

We believe both the IT and energy industries can benefit tremendously from a deep understanding of data center dynamics as a cyber-physical system and can contribute to operating them efficiently at low cost. Modeling and simulation environments that allow users to quickly explore "what-if" scenarios can reduce the risk of adopting alternative and renewable energy sources in the IT industry. Software techniques, such as replication and fault tolerance can shed new lights in energy efficiency. Exploring geo-distribution, workload redirection and migration may significantly reduce the cost of building data centers. Recovering and reusing waste heat generated by data centers can reduce the overall burden on energy grids. The challenge is significant and the task requires joint efforts from many disciplines, across industry, government, and academia.

References

[1] Barroso, L. A., & Hölzle, U. (2009). The datacenter as a computer: An introduction to the design of warehouse-scale machines. Synthesis Lectures on Computer Architecture.

[2] Sankar, S., & Vaid, K. (2009, October). Storage characterization for unstructured data in online services applications. In Workload Characterization, 2009. IISWC 2009. IEEE International Symposium on (pp. 148-157). IEEE.

[3] Delimitrou, C., Sankar, S., Vaid, K., & Kozyrakis, C. (2011, November). Decoupling datacenter studies from access to large-scale applications: A modeling approach for storage workloads. In Workload Characterization (IISWC), 2011 IEEE International Symposium on (pp. 51-60). IEEE.

[4] Gupta, S. K., Gilbert, R. R., Banerjee, A., Abbasi, Z., Mukherjee, T., & Varsamopoulos, G. (2011, July). GDCSim: A tool for analyzing green data center design and resource management techniques. In Green Computing Conference and Workshops (IGCC), 2011 International (pp. 1-8). IEEE.

[5] Herrero-Lopez, S., Williams, J. R., & Sanchez, A. (2011, September). Large-scale simulator for global data infrastructure optimization. In Cluster Computing (CLUSTER), 2011 IEEE International Conference on (pp. 54-64). IEEE

[6] Lim, S. H., Sharma, B., Nam, G., Kim, E. K., & Das, C. R. (2009, August). MDCSim: A multi-tier data center simulation, platform. In Cluster Computing and Workshops, 2009. CLUSTER'09. IEEE International Conference on (pp. 1-9). IEEE

[7] Meisner, D., Wu, J., & Wenisch, T. F. (2012, April). BigHouse: A simulation infrastructure for data center systems. In Performance Analysis of Systems and Software (ISPASS), 2012 IEEE International Symposium on (pp. 35-45). IEEE

[8] Miller, R. “Inside Microsoft’s Chicago Data Center”. Datacenter Knowledge, Oct 2009