Cooperative Expendable Micro-Slice Servers (CEMS): Low Cost, Low Power Servers for Internet-Scale Services

James Hamilton
Amazon Web Services
1200 12th Ave. S.
Seattle, WA 98144
http://mvdirona.com/jrh/work
James@amazon.com

ABSTRACT

The Cooperative Expendable Micro-Slice Servers (CEMS) project evaluates low cost, low power servers for high-scale internet-services using commodity, client-side components. It is a follow-on project to the 2007 CIDR paper Architecture for Modular Data Centers [11]. The goals of the CEMS project are to establish that low-cost, low-power servers produce better price/performance and better power/performance than current purpose-built servers. In addition, we aim to establish the viability and efficiency of a fail-in-place model. We use work done per dollar and work done per joule as measures of server efficiency and show that more, lower-power servers produce the same aggregate throughput much more cost effectively and we use measured performance results from a large, consumer internet service to argue this point.

Categories and Subject Descriptors

B.8.8.m [Hardware PERFORMANCE AND RELIABILITY Miscellaneous]: Server design; low-power, low-cost client and embedded parts in servers.

General Terms

Measurement, Performance, Economics, Experimentation.

Keywords

Data centers, power efficiency, low cost servers, low power servers, total cost of ownership.

1. Introduction

In this paper we make the following points: low cost, low power servers in aggregate can produce the same throughput as conventional purpose-built servers at lower initial hardware cost (the number one cost in an internet service), and at lower operational cost (cooling, power provisioning, and power are the most significant of these). We include detailed descriptions of the prototype server we built to gather performance data using a high-scale, consumer internet service and show that our prototype server can produce 56% the throughput at 21% the cost and 13% the power. We'll present the case that low cost, low power systems are more cost effective than purpose-built servers, and outline future work planned to show that 1) the fail-in-place, service-free model is actually cost effective and 2) the increased server-mortality rates driven by using lower cost, lower quality parts does not negatively impact their price/performance advantage.

Power has become the most important issue for high-scale data center operators in the past two years [3]. This is driven by cost and social issues. The popular press, congress and the EPA are all concerned about data center power consumption nation-wide [6]. To address the high–scale data center power problem, we first need to understand where the power is used and let that result set the direction for CEMS project.

This work follows from the earlier work, Architecture for Modular Data Center [11], where we are argued that data centers are better constructed incrementally from prefabricated modular components rather than the current industry practice of building from large monolithic designs. The modular data centers work achieved improvements in cooling system efficiency, allowed incremental growth that more closely track internet service growth requirements, and moved much of the infrastructure from a 15-year cycle of innovation to a 3-year cycle. Earlier this year it was publically announced that the first large-scale, commercial modular data center deployment will come on line in early 2009 [12].

The project reported in this paper first investigates where the costs are in a large-scale data center and concludes that the dominant costs are 1) servers and 2) costs functionally related to power. We then look at power in more detail to understand where the power is dissipated in a high-scale service. Understanding that server costs and costs functionally related to power dominate, we then propose a new server design optimizing for these two factors and compare this new design with a commercial server using a high-scale service production workload.

2. H/W and Fully Burdened Power Dominate

Before attempting to reduce high-scale data center costs, we need to understand the most significant costs in more detail. Analysts and the popular press often report that power is the single largest cost in high-scale data centers. This isn't entirely true as stated, but power is clearly is one of the fastest growing costs [3]. Other

4th *Biennial Conference on Innovative Data Systems Research (CIDR)*
January 4-7, 2009, Asilomar, California, USA.

reports suggest that people costs dominate [5]. People costs often are dominant in enterprise data centers, however, in high-scale facilities with tens of thousands of servers, server administration is heavily automated [10] and, once it has been, administration costs fall below 10% and often below 5%.

In order of magnitude from largest first, the most significant costs are 1) server acquisition, 2) cooling, 3) power distribution, and 4) power itself.

To understand this data in more detail, we model a $200M facility capable of delivering 15MW of critical load (server power) with the following assumptions:

- Facility: ~$200M for 15MW DC (15 yr Amortization)
- Servers: ~$2k/each, roughly 50,000 (3 yr Amortization)
- Commercial Power: ~$0.07/kWh
- 5% cost of money

To compare these costs, we need to normalize long lived capital costs with 15-year amortization periods and short lived capital costs having 3-year amortization periods. In addition we need to compare monthly operational costs with these capital costs in order to be able to understand which are the most important. We normalize by assuming a 5% annual cost of money with monthly payments and essentially borrow the money for capital expenses and pay back monthly over the assumed life and amortization period of the equipment. This converts the capital expenses to effective cost per month. And, by considering amortization periods, we normalize long lived and short lived capital and recognize each appropriately. In this model, land, taxes, security and administration are not included due to their relatively small contribution to overall costs.

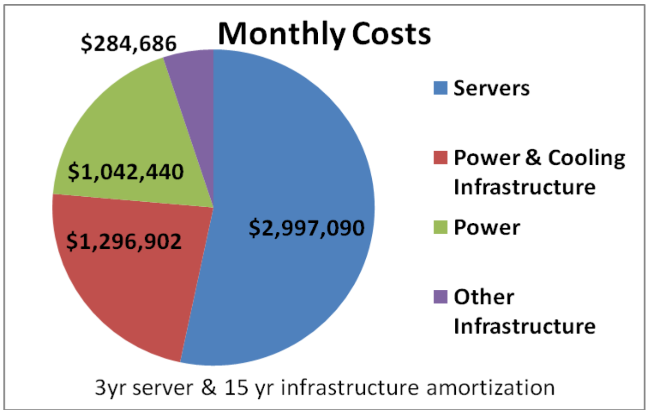

Figure 1: Monthly Server, Power, and Infrastructure Costs

Figure 1 shows that power costs are much lower than infrastructure costs, and also much less than the servers themselves. Servers are the dominant cost, but, before we conclude that power is only 23% of the total, it's worth looking more closely. Infrastructure includes the building, power distribution, and cooling. Power distribution and cooling make up 82% of the costs of infrastructure [2] with the building itself down in the 12-15% range. Power distribution is functionally related to the power consumed in that sufficient power distribution equipment is required to distribute the maximum amount of power consumed. Cooling is also functionally related to power in that the heat from all power that is dissipated in the building must be removed. The vast majority of the infrastructure cost is functionally related to power. We define the fully burdened cost of power to be the sum of power, power distribution, and cooling costs.

3. Where Does the Power Go?

To get started, it helps to define a few terms. *Total Facility Power* is the power delivered by the utility to the property line of the data center. *IT Equipment Power* is the power delivered to the critical load, the servers in the data center. The difference between Total Facility Power and IT Equipment Power is the power lost in power distribution and in cooling the facility. Effectively, this difference is the facility infrastructure overhead.

The Green Grid defines two useful terms when looking at data center efficiency: *Power Usage Effectiveness* (PUE) and *Data Center Infrastructure Effectiveness* (DCiE) [9].

PUE = (Total Power) / (IT Equip. Power)

DCiE = (IT Equip. Power) / (Total Power) * 100%

Power Usage Effectiveness is Total Facility Power over IT Equipment power. The PUE tells us how many watts must be delivered to the data center in order to get one watt to the critical load, the servers themselves. DCiE is the reciprocal of PUE and is defined as IT Equipment Power over Total facility power. DCiE tells us what percentage of the power delivered to the facility actually gets delivered to the servers.

These terms both have the same information content. A PUE of 1.7 states that for every watt delivered to the IT equipment (the servers), we dissipate 0.7W in power distribution and mechanical systems (air conditioning, pumps, fans, etc.). A PUE of 1.7 is the same as a DCiE of 59% which states that for every watt delivered to the facility, 59% is delivered to the IT equipment. The DCiE also tells us that 41% of the power delivered to the data center is lost in power distribution and cooling overhead.

PUEs vary greatly. Very inefficient enterprise facilities are often as low as 2.0 or even 3.0 [9] and unusual, industry-leading facilities are being advertised as better than 1.2 [8]. These latter reports, however, are difficult to corroborate.

In this exploration into power losses, we'll consider a current-generation facility. This is one that would be built if current, well understood techniques are applied and good quality but widely available equipment is deployed. This test facility has a PUE of 1.7, putting it much better than most of the world's data centers. But it is not using some of the latest, not yet well documented innovations. A PUE of 1.7 is far above average but lower than the best and forms a good baseline for us to look at to understand where the power is going and where the largest inefficiencies lie.

Looking more deeply at our PUE 1.7 facility, we know by the definition of PUE that we are delivering 59% of the data center power to the IT equipment. We need to understand where the remaining 41% is going.

To understand where the 41% lost to data center infrastructure is going, we look first to the power distribution equipment since it is both easier to inventory and these distribution loses are easier to track. Looking at figure 2, we can see every conversion and the efficiency of each conversation from the power delivered by the utility at 115,000V through to deliver to the servers at 208V.

Starting at the upper left corner of Figure 2, we see the utility delivers us 115kV and we first step it down to 13.2kv. The 13.2kv feed is delivered to the Uninterruptable Power Supply (UPS). In this case we use a battery-based UPS system, but rotary systems are also common. This particular battery-based UPS is 94% efficient, taking all current through rectifiers to direct current and then inverting it all back to AC. Rotary designs are usually more efficient than the example shown here and bypass designs can exceed 97% efficiency. In this example, a non-bypass UPS installation, all power flowing to UPS protected equipment (the servers and most of the mechanical systems) is first rectified to DC and then inverted back to AC. All the power destined to the servers flows through these two conversions steps whether or not there is a power failure, and these two conversion steps contribute the bulk of the losses, bringing down the UPS efficiency to 94%. More efficient bypass UPSs avoid these losses by routing most power "around" the UPS in the common, non-power failure case.

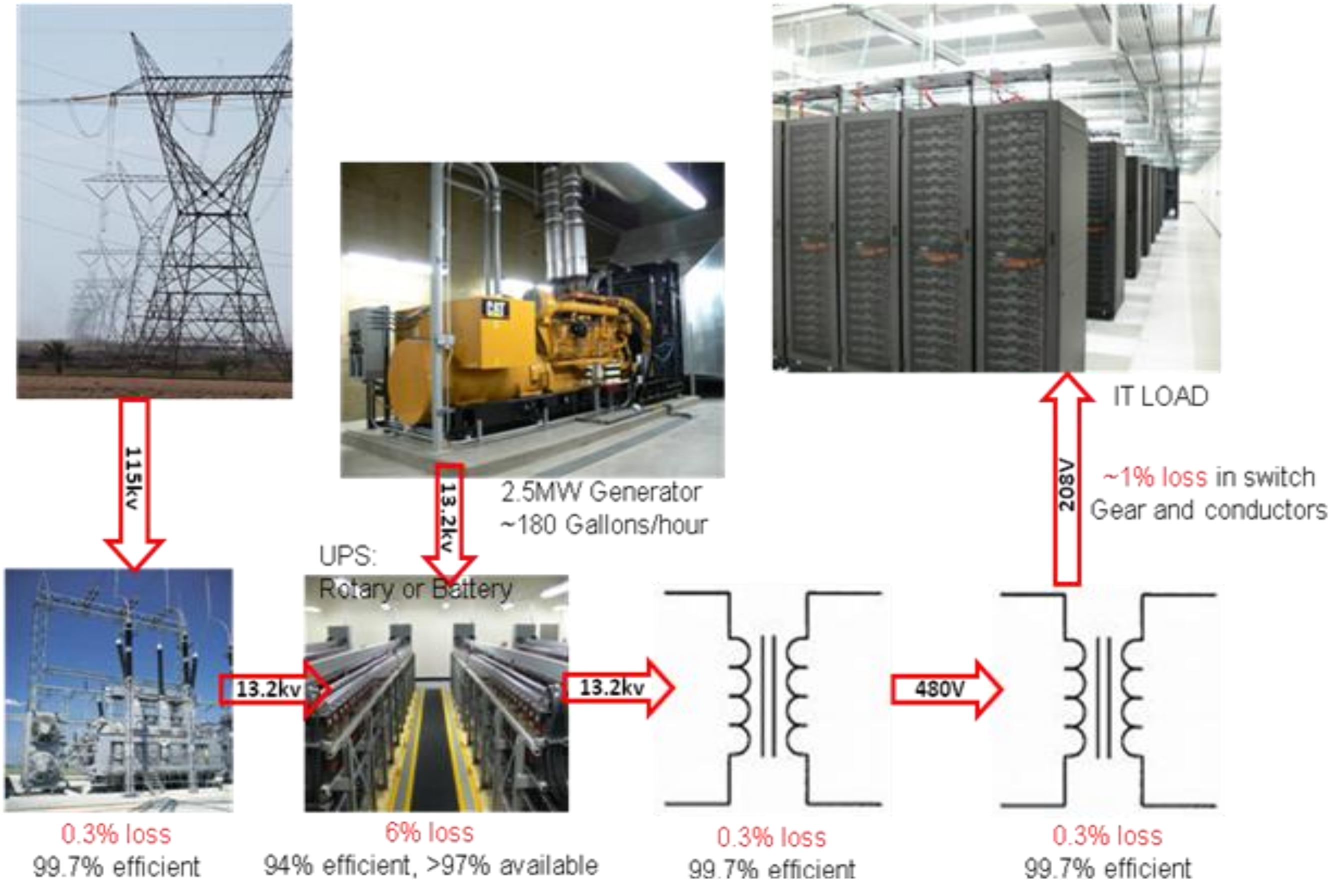

Figure 2: Power Distribution

For longer term power outages, there are usually generators to keep the facility operational. The generation system introduces essentially no additional losses when not being used but they greatly increase the capital expense with a 2.5MW generator pricing out at more than $2M. Most facilities will have at least 1 extra generator (N+1) and many facilities will have 2 spares (N+2) allowing one to be in maintenance, one to fail on startup and still to be able to run the facility at full load during a power failure. A 2.5MW generator will burn just under 180 gallons/hour of diesel so environmentally conscious operators work hard to minimize their generator time. And the storage of well over 100,000 gallons of diesel at the facility brings additional cost, storage space, insurance risk, and maintenance issues.

After the UPS, we step down the 13.2kV voltage to 480V and then that is further stepped down to 208V for distribution to the critical load, the servers. In this facility, we are using very high quality transformers, so we experience losses of only 0.3% at each transformer. We estimate that we lose a further 1% in switch gear and conductor losses throughout the facility.

In summary, we have three 99.7% efficient transformers, a 94% efficient UPS and 1% losses in distribution for an overall power distribution loss of 8% (0.997^3*0.94*0.99 => 0.922).

We know we deliver 59% of the facility power to the critical load and, from the electrical distribution system analysis above, we know we lose 8% of total power to power distribution losses. By subtraction, we have 33% lost to mechanical systems responsible for data center cooling.

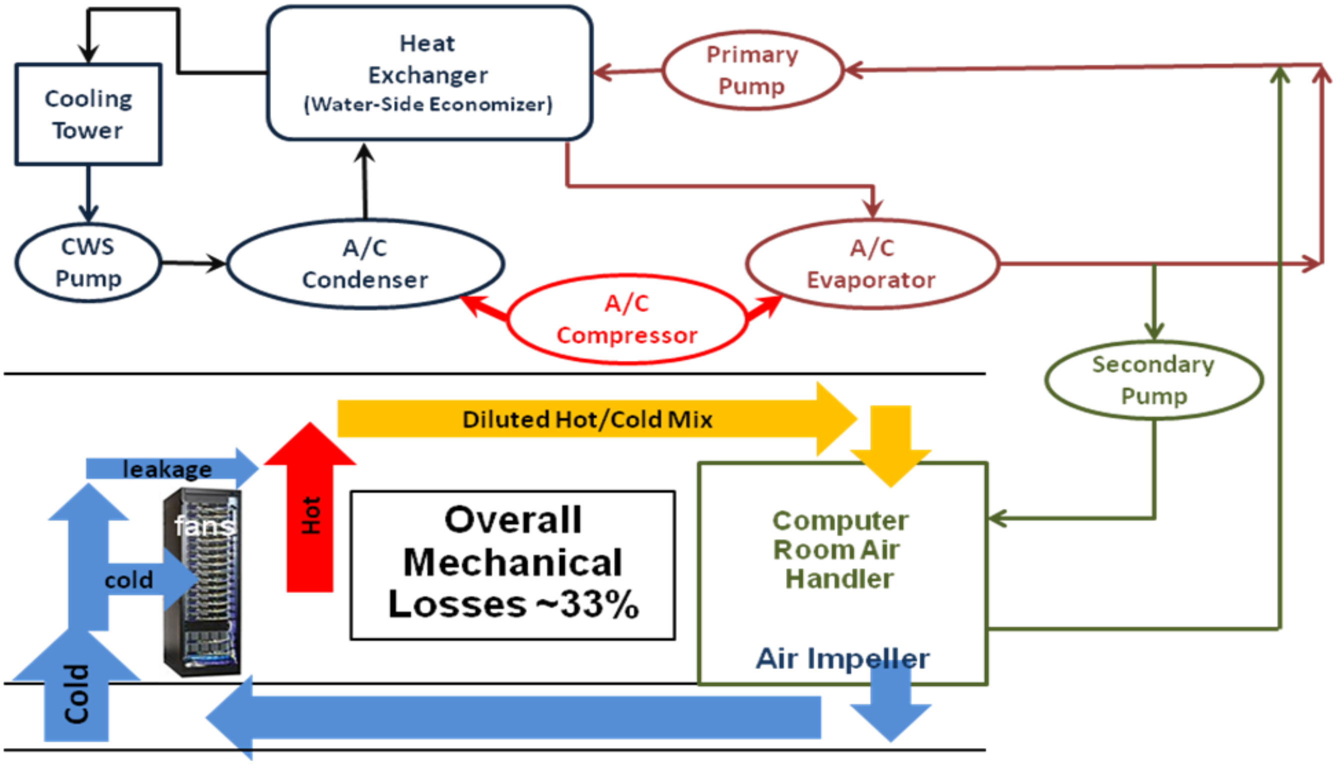

Figure 3: Mechanical Systems

Several observations emerge from this summary. The first is that power distribution is already fairly efficient. Taking the 8% efficiency number down to 4% to 5% by using a 97% efficient UPS and eliminating 1 layer of power conversion is an easy improvement. Further reductions in power distribution losses are possible but the positive impact can't exceed 8% so we're better rewarded looking to improvements in the mechanical systems, where we are spending 33% of the power, and in the servers, where we are dissipating 59% of the power.

The CEMS project focused on the latter, increasing the efficiency of the servers themselves.

4. CEMS Introduction

From the previous section, we understand that 59% of the power dissipated in a high-scale data center is delivered to the critical load. Generally that is a good thing in that power delivered to the servers is success from a data center infrastructure perspective. All power delivered to the server has a *chance* of actually getting work done. However, with the bulk of the power going to servers, server efficiency and utilization clearly will have a substantial impact on overall data center power efficiency. In this work, we focus on the former, server efficiency.

The CEMS project originated from two core observations: 1) nearly 60% of the power delivered to a high-scale data center is delivered to servers, so server efficiency has a dominant impact on overall system (data center and server) efficiency, and 2) newer servers design are increasingly out of balance as CPU performance increases without matching improvements in memory and storage subsystems. Let's look first at the balance issue in more detail and then come back to how to leverage these two observations to deliver a reliable service while substantial lowering costs and increasing power utilization efficiency.

4.1 System Balance

Looking back 25 years, we have experienced steady improvement in CPU performance and, for a given algorithm, increased performance generally requires increased data rates. In the high performance computing world, this is reported in bytes/FLOP but it's just as relevant in the commercial processing world. More CPU performance requires more memory bandwidth to get value from that increase in performance. Otherwise, the faster processor just spends more time in memory stalls and doesn't actually get more work done. For the bulk of the last 25 years, CPU performance improvements have been driven by design improvements and clock frequency increases. Having hit the power wall, we're now less reliant on clock frequency improvements than in the past and more dependent upon increases in core counts. But the net is that processor performance continues to grow unabated and this is expected to continue.

Looking at the first row of Table 1, from Dave Patterson's "Latency Lags Bandwidth" paper [16], we can quantify the argument above. The data in Table 1 are extracted from leading commodity components over the last 25 years and what is reported is the multiplicative performance increase per year. Looking at this chart, we see that CPU bandwidth is growing at 1.5x per year whereas memory bandwidth, LAN bandwidth, and disk bandwidth are all growing more slowly.

Table 1. Annual Bandwidth and Latency Improvements

Annual Improvement	CPU	DRAM	LAN	Disk
Bandwidth	1.5	1.3	1.4	1.3
Latency	1.2	1.1	1.1	1.1

The core argument in the Patterson paper is that latency is an even bigger problem in that latency is driven by physical limits whereas bandwidth can be addressed through parallelism. Essentially it is always possible to get more bandwidth by adding more communications paths between CPU and memory but this consumes more power and drives up costs and processor pin counts are difficult to continue to grow by factors. This general problem of CPU bandwidth and latency improvements outstripping those of memory are often referred to as the memory wall.

It’s clear that the memory wall really is a problem but it’s equally clear that we have at least two broad alternatives in addressing the problem: 1) we can invest in higher bandwidth communications between the processors and memory or 2) we can avoid the problem by using more servers with cheaper, lower-powered processors that more closely match the capabilities of the memory subsystem. In short, we can invest in fixing the problem or we can chose to defer the issue by using cheaper, lower powered CPUs with lower bandwidth requirements.

Looking at the first solution, improving memory-to-CPU bandwidth, many potential solutions exist but all have compromises. The simplest is to add more memory channels and let parallelism get us the bandwidth we need. This simple solution will have positive impact, but adding package pins drives up costs and power consumption. Other solutions with great potential are on-chip optical [7] and chip stacking [4]. On-chip optical has great promise, but commercial application of this technology is likely 10 to 15 years away so it’s not a short-term solution. Chip stacking and using Through Silicon Vias (TSV) as chip interconnections is a nearer term solution that has achieved commercial use in embedded applications and is expected to be applied to servers in the near future.

We have no doubt that both these techniques will be employed and will have positive impact on this problem over time. What we look at more closely in this project is avoiding the problem entirely or, more accurately, deferring the problem by using lower-powered, higher volume, cheaper processors in greater numbers.

Given that large service workloads are already partitioned and running over 10^2+ servers, there is an opportunity to use more, less powerful servers to support the same workload. From Section 2, Hardware and Fully Burdened Power Dominates, we know that high-scale service costs models are dominated by hardware and fully burdened power costs. From section 4.1, System Balance, we know that CPU bandwidth consumption is outstripping memory bandwidth and servers are getting increasingly out of balance. All these factors indicate that using greater numbers of high volume, lower performance, lower power parts will have the aggregate performance needed to support the workload. Our goal with CEMS is to investigate the practicality of this design point using a production, high-scale data center workload. We’ll investigate the practicality of low power designs and the savings in power, purchase cost, and run a long term study to understand server mortality losses due to the use of lower quality, non-server targeted hardware components.

4.2 CEMS Design

The arguments up to this point suggest that there is opportunity to change how we are building servers for high-scale data centers. Given that hardware and costs functionally related to power dominate the overall cost of operations for high-scale services (Section 2), it is clear we should be making server design decisions on the basis of work done per dollar and work done per joule. Existing server designs are lower volume than client and embedded parts, and therefore more expensive. Existing server designs, unlike embedded and mobile clients, do focus on system performance, but this performance comes at an increased power and purchase cost. Existing servers are designed to run reliably 24x7 for years without failure, but this additional quality also comes at a cost. Many servers are replaced after 3 years and most are replaced before 5 years, due to older systems being less power efficient [13]. In Section 4.3, Performance Results, we will show that, at least in some cases, purpose-built server designs can’t do as much work within the existing power envelope.

When we replace servers well before they fail, we are effectively paying for quality that we’re not using. High-scale services can continue to achieve their SLA commitments in the presence of server failure. In fact, at any fixed point in time it’s rare not to have 3% to 5% of a large server farm unavailable for customer workload. Understanding that individual server failures in isolation don’t negatively impact the service, over-engineering server quality to avoid failure brings additional cost without delivering additional value. Ideally we should be maximizing work done per dollar, work done per joule, and servers should be failing approximately when they are scheduled for replacement.

The CEMS project investigates a different design point on the basis of the arguments above. Summarizing the observations above: 1) servers are increasingly out of balance, with CPU bandwidth increasing much faster than memory bandwidth, 2) servers are engineered to last long periods of time and yet are replaced in 3-5 years, 3) servers emphasize performance rather than optimizing for work done per cost unit and work done per energy unit, and yet server costs and energy consumption dominate overall service costs, and 4) client and embedded parts volumes are several orders of magnitude higher than servers, and consequently are less expensive.

The CEMS projects investigates if high-scale, commercial services can be operated more efficiently using low-cost, low-power client or embedded components. And, do the increased failure rates from using non-server components in 24x7, high-load operation increase costs beyond the original savings at purchase time?

Our desire to test the systems using production, commercial service workloads constrained the server design we adopted. We accepted these constraints because we felt it was much more relevant to understand how this server design performed on production workloads rather than benchmarks. And, we wanted to be able to compare their performance to existing, purpose-built servers from a major server supplier. Real workloads are more interesting and more credible than benchmarks but real workloads are large and difficult to rewrite. Accepting this constraint, we designed the hardware to be capable of running Windows Server 2003 and the existing application implementation unchanged, with the same disk and memory configuration. This allows us to quickly test the new server design, but also restricts the gains possible from the low-cost, low-power approach. It’s a data point to show the design approach works and will allow us to run long-term reliability tests on real workloads in production. We’ll investigate other design possibilities and directions in the future work section.

The service we selected was chosen because it is one of the larger services at Microsoft, with thousands of identical servers, and their IIS-based workload has broad application across the industry. Essentially, they are typical of a broad equivalence-class of services. Finally, they are profitable and, by extension, have at least reasonable control of their costs of operation. Their server SKU selection was made carefully to be efficient for their service and it was supplied by one of the larger server providers. Their existing servers are 3.6Ghz processors with two small enterprise class disks (15k RPM SCSI) and 2GB of main memory. These servers draw 407W at full load but, to increase operational stability and to allow usage spike headroom, the service aims to run them at 60% of full load. At 60% load, these existing servers draw 297W.

The CEMS design aims to support the same workload without change to the workload or to the operational characteristics (it will also be run at 60% load). With these constraints, the goal is to get at least 2x more work done per unit cost and at least 2x more work done per unit power. As a design partner we selected Rackable Systems [17]. Rackable was the first company commercially shipping modular data centers and focuses exclusively on server-side computing. The design we selected:

- AMD Athlon 64 x2 4850e at 2.5Ghz
- 2G 2 x 2gb DDR2 533mhz unbufferd ECC 3.2W
- I-BASE MI930 Min-ITX ATi M690T / SB600 chipset
- Custom Rackable Sled chassis

To further reduce costs and to improve cooling efficiency, the rack design is based upon placing the systems on sleds rather than the more standard, fully enclosed server chassis. The sleds are just strong enough to support the servers, disk, and power supplies. The sled approach has the advantage of reducing materials costs, slightly lower manufacturing costs, reduced resistance to cooling airflow and substantially reducing servicing costs.

We use a single, low-power 2 ½" disk per server for compatibility with the existing hardware, which also used 1 logical disk drive. When we can justify software changes, we'll update this H/W design to a single shared disk per 6-server sled. Another design we have not fully investigated but appears to have merit is to use a single or small number of flash memory-based SSDs at the rack level supporting boot, audit logging, and diagnostic logging for all 240 servers in the rack. Currently, the single disk per server increases costs unnecessarily by about 10% over shared disk designs. However, it's simple and the efficiency loss doesn't appear to obscure the price/performance and power/performance benefits of the CEMS design.

Each sled houses six servers, six 2.5" disks and a single shared power supply. As mentioned above, we plan to move to a single shared disk but, for simplicity, the current approach has one disk per server. In the current design, we are running individual network cables from the top of rack switch to each server. In the higher volume design going into test, we plan to have one 8-port mini-switch per sled. Using a $50 8-port mini-switch for every six-server sled allows us to reduce cabling cost and complexity and to reduce the required port count on the top of rack switch, making the mini-switch an overall cost savings.

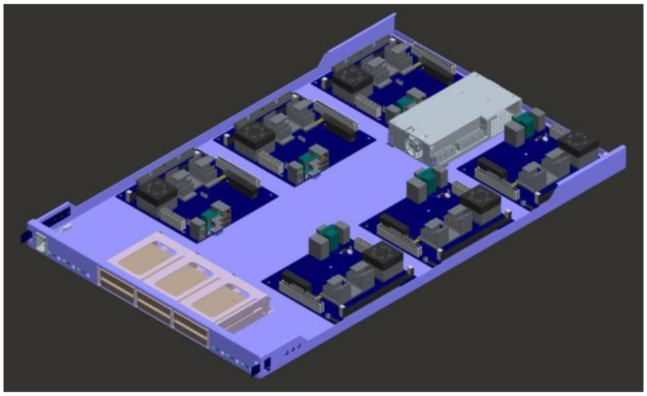

Figure 4: 6-Server, 1U (1 Rack Unit) CEMS Sled

Figure 4 shows that all six servers on a sled share a power supply. This gives us cost and potential efficiency advantages, but means that a power supply failure will bring down six servers instead of one. Looking at long term failure rates on services on which we have worked, we see that failures are dominated by disk and memory failures, so we decided to save costs by sharing a power supply. With thousands of servers in the farm, the increase in correlated failures caused by a shared power supplies is believed to be a minor factor.

4.3 Performance Results

The high-scale, commercial internet service we're partnered with in this study investigates new hardware SKUs by loading their application code on the system under test and running a simulated production workload against it measuring the number of requests serviced per second achieved at 60% CPU load. The CPU load is capped at 60% load, since this is the load they target in production to obtain traffic handling headroom and increased service stability. The workload under test is a IIS web server workload with minimal disk activity characteristic of the middle tier of many high-scale services.

In these tests we compare the existing hardware used by the service, labeled System-X, with CEMS. All tests are run with the same production qualified application code running under unmodified Windows Server 2003. In each test the Request/Second (RPS) rate is increased slowly until CPU load as measured by Windows performance counters was just less than 60%. At just less than 60% CPU load, the request per second rate (RPS) and overall server power draw are both recorded.

Through earlier investigations with early CEMS prototypes (V1 and V2 in Table 2), the workload appears to scale fairly linearly with clock frequency and with core counts. Understanding that and wanting to minimize power consumption for a given amount of work done argues for using many, low power cores. For the final CEMS design we will take into full scale production early next year to do long term failure analysis, we elected to use a 2-core desktop part, the AMD Athlon 4580e operating at 2.5Ghz. In Table 2 below, we compare the performance of the existing hardware, System-X, with the current Athlon 4580e CEMS design, and the previous CEMS generations based upon the Athlon 3400e and the 2000+ respectively. The early CEMS parts were much lower power at less 1/3 of the final CEMS server but, the important metric, work done/joule was worse as was work done per dollar.

Table 2. CEMS Performance Results

Compare	Sys-X	CEMS V3 (Athlon 4850e)	CEMS V2 (Athlon 3400e)	CEMS V1 (Athlon 2000+)
CPU %	56%	57%	57%	61%
RPS	96.0	75.3	54.3	17.0
Price	$2,371	$500	$685	$500
Power	295J	60J	39J	33J
RPS/$	0.04	0.15	0.08	0.03
RPS/Joule	0.33	1.25	1.39	0.52
RPS/Rack	1,918	18,062	13,025	4,080

The current server, designated System-X in Table 2, is broadly deployed by this service with thousands of servers in production. Comparing CEMS V3 with System-X, we are not surprised to see the CEMS server delivering lower throughput than System-X. Our design point is work done/price and work done/joule rather than raw performance, so 78% (75.3/96.0) throughput is not by itself a problem.

In what follows, we refer to CEMS V3 as simply CEMS. We measure work done in application Requests/Second (RPS). Looking at RPS/dollar, we see quite favorable results with the CEMS server producing 3.7x the RPS/dollar of System-X. The lower individual server throughput is more than offset by much lower cost. System-X is $2,371/server whereas CEMS is $500/server, giving substantially better value as long as failure rates aren't significantly higher. CEMS is a 375% (0.15/0.04) better price/performer than System-X.

4.3.1 RPS/Joule

A joule is a watt second. What we're measuring in this comparison is the number of requests that can be serviced in a watt second. Again we see the lower individual server performance is more than offset by a significant reduction in power consumption. System-X processed 96 requests using 295 joules or 0.33 requests/joule. The CEMS server processed 75 requests in 60 joules or 1.25 requests/joule. CEMS is a 379% (1.25/0.33) better performer/joule than System-X.

4.3.2 RPS/Rack

In *Why Blade Servers are not the Answer to all Questions* [14], we argue that gratuitous server density – density without value – is a bad idea and performance/dollar and performance/joule are better optimization points. Nonetheless, we include it here in recognition that performance density does drive many purchasing decisions. And there are locations such as Hong Kong and New York where server density can be important. CEMS is a 942% (18,062/1,918) better performer/rack than System-X.

In summary, CEMS exceeds System-X by fairly substantial margins:

- RPS/dollar:3.7x
- RPS/Joule: 3.9x
- RPS/Rack: 9.4x

5. Future Directions

The system supplier of System-X has produced a follow-on server of System-X we'll refer to as System-X' here. Unfortunately none of the System-X' servers have been installed at the service so we were unable to get a solid performance measurement. But from reading the specifications on System-X', they appear to have taken a design approach very similar to that of CEMS, apparently optimizing for work done/dollar and work done/joule, in that we see substantially improved price, power, and density improvements but we don't expect substantially changed performance based upon early integer micro-benchmark runs. In our view, this is exactly where the industry should be going and it's good to see.

Without a system to measure, it's impossible to know with certainty the performance of System-X' but its estimated to be nearly a factor of two better than System-X in RPS/dollar, RPS/Joule, and RPS/rack but roughly the same as System-X in raw performance (RPS). If that data is accurate, CEMS will continue to have nearly a 2x advantage across our three of our dimensions of interest.

System-X' is not expected to upgraded again for another 12 to 18 months and so this lead is likely stable. In addition, we see opportunity to improve CEMS RPS/joule by upwards of 50% using Intel Atom or unannounced components from AMD. We see opportunity to improve CEMS pricing by eliminating the dedicated disk/server and going to a single disk/sled (six servers) design. It's not been fully investigated, but we may also be able to go with a single SSD per rack, eliminating the power draw and cost of 240 disks.

The workload hosted by CEMS in this work is a fairly typical web server workload with minimal disk activity. Can we apply the same approach of using redundant, client-side component in the storage tier?

In the next phase of CEMS testing, we will put a rack of CEMS into production beside hundreds of System-X and a small number of System-X' racks to study long term software failure rates, hardware failure rates, and overall cost of ownership differences between System-X and CEMS. We will also use results from the long term full rack testing program to investigate the fail-in-place, service-free model suggested by [11] at CIDR2007.

6. Related Work

The move away from expensive, mainframe class scale-up servers to commodity, scale-out servers has been underway for over a decade. But these scale-out servers now widely deployed in support of internet-scale workloads are still high-quality, purpose-built designs. We propose instead the use of client-side and embedded components in server design.

It may have been David Patterson who first observed that embedded design techniques are beginning to have substantial overlap with server design as power becomes the dominant factor in each. I first saw this observation documented in [1], on which Patterson was a co-author. I agree completely, and that observation has influenced this work.

At ISCA earlier this year, Lim, Ranganathan, Chang, Patel, Mudge, and Reinhardt proposed a benchmark suite for warehouse computing workloads, and a server design based upon non-server components [15]. This is excellent work in that they 1) build actual hardware prototypes, and 2) evaluate them with respect to work done per dollar.

Our work here is similar to Lim et al. in targeting embedded and client-side components in the design of servers and in producing operational hardware prototypes. But our metrics of interest are somewhat broader, looking at work done/dollar, work done/joule, and work done/rack and we use an actual internet-scale workload rather than a synthetic benchmark.

7. Conclusions

In this work, we focus on establishing that work done/dollar and work done/joule are the correct measures of server value for high-scale services. We show that hardware costs and fully burdened power costs dominate the cost of delivering high-scale services. We report on investigations into where the power is dissipated in a high-scale data center. This investigation serves two purposes: 1) it shows where more research into power savings can deliver value, and 2) it provides the motivation for our work on Collaborative Expendable Micro-slice Servers (CEMS).

We documented the CEMS server prototype hardware design and showed performance results of CEMS running a production, commercial service workload. We compared these performance findings against the server design currently in use by this commercial service and showed the new design is superior when measured by work done/dollar, work done/joule, and work done/rack. Finally we discuss future plans to improve the CEMS design and further drop power consumption by 50% and reduce cost further without negatively impacting the metrics of interest.

Our findings support the assertion that current server designs are not well optimized for mega-services and that custom power and cost optimized server designs can produce much better value.

8. Acknowledgements

This work was initiated while the author was a member of the Microsoft Live Platform Services team and continued when the author transferred to the Microsoft Data Center Futures team.

We would like to thank David Treadwell (Microsoft Live Platform Services) for early support of this work and the Microsoft Data Center Futures team for support through the design phase and production workload tests. Albert Greenberg of Microsoft Research contributed throughout.

This project would not have happened without the deep involvement, support, and contribution of Giovanni Coglitore and the entire Rackable Systems team (http://www.rackable.com/). Rackable Systems built the prototypes and drove many of the design decisions. Jennifer Hamilton of Microsoft contributed substantially as did Joe Hellerstein of Berkeley. Also instrumental in advancing this work were Phil Smoot, Ed Fancher, David Schiffrin, Rajeev Shukla, Glenn Beeswanger, Eliot Gillum, Suresh Babu, Cindy Sherrett, Amaya Souarez, and Kushagra Vaid, all of Microsoft.

9. References

[1] K. Asanovic, R. Bodik, et al. The Landscape of Parallel Computing: A View From Berkeley, http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.pdf, Dec. 2006.

[2] C. Belady & M. Manos, Intense Computing or In Tents Computing. http://blogs.msdn.com/the_power_of_software/archive/2008/09/19/intense-computing-or-in-tents-computing.aspx, Sept. 2008.

[3] K. Brill. The Invisible Crisis in the Data Center: The Economic Meltdown of Moore's Law. Uptime Institute White Paper, http://uptimeinstitute.org/index2.php?option=com_docman&task=doc_view&gid=22&Itemid=162, 2007.

[4] B. Dang, et al. 3D Chip Stacking with C4 Technology. *IBM Journal of Research and Development*, Oct. 2008.

[5] V. Ercegovac, J. Glider, et al. Impliance: An Information Management Appliance. *Conference on Innovative Data Systems Research*, http://www-db.cs.wisc.edu/cidr/cidr2007/slides/p41-lohman.ppt, Jan. 2007.

[6] EPA. EPA Report to Congress on Data Center Energy Efficiency, http://www.energystar.gov/index.cfm?c=prod_development.server_efficiency#epa, Aug. 2007.

[7] S. Firth. (Tiny) Rings of Fire, HP Labs, http://www.hpl.hp.com/news/2008/oct-dec/photonics2.html, Dec. 2008.

[8] Google. Commitment to Sustainable Computing, http://www.google.com/corporate/data centers/, Oct. 2008.

[9] Green Grid. The Green Grid Power Efficiency Metrics: PUE & DCiE, http://www.thegreengrid.org/gg_content/TGG_Data_Center_Power_Efficiency_Metrics_PUE_and_DCiE.pdf, 2007.

[10] J. Hamilton. On Designing and Deploying Internet-Scale Services, *USENIX Large Installation Systems Administrators Conference*, Nov. 2007.

[11] J. Hamilton. Architecture for Modular Data Centers, Conference on Innovative Data Systems Research 2007, Jan. 2007.

[12] J. Hamilton. First Containerized Data Center Announcement, http://perspectives.mvdirona.com/2008/04/02/FirstContainerizedData centerAnnouncement.aspx, Apr. 2008.

[13] J. Hamilton. Annualized Fully Burdened Cost of Power, http://perspectives.mvdirona.com/2008/12/06/AnnualFullyBurdenedCostOfPower.aspx, Dec. 2008.

[14] J. Hamilton.Why Blade Servers are not the Answer to all Questions, http://perspectives.mvdirona.com/2008/09/11/WhyBladeServersArentTheAnswerToAllQuestions.aspx, Sep. 2008.

[15] K. Lim, P. Ranganatham, et al. Understanding and Designing New Server Architectures for Emerging Warehouse-Computing Environments. *International Symposium on Computer Architecture*, 2008.

[16] D. A. Patterson. Latency Lags Bandwidth. *Communications of the ACM*, 47(10), 2004.

[17] Rackable Systems, http://www.rackable.com/.