QPS,KW-hr,MTBF,ΔT,PUE,IOPS,DB/RH: A Day in a Life of a Datacenter Architect

Kushagra Vaid
Principal Architect, Datacenter Infrastructure
Microsoft Online Services Division
kvaid@microsoft.com
First... lets explain the lingo in the talk title...

<table>
<thead>
<tr>
<th>Metric</th>
<th>Definition</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>QPS</td>
<td>Queries per second</td>
<td>Performance of Index-serving engine</td>
</tr>
<tr>
<td>KW-hr</td>
<td>Kilowatt-hr: Energy consumption metric</td>
<td>Effectiveness of power conservation schemes</td>
</tr>
<tr>
<td>MTBF</td>
<td>Mean Time Between Failures</td>
<td>Understanding component reliability</td>
</tr>
<tr>
<td>ΔT</td>
<td>Temperature delta between front and rear of server</td>
<td>Server chassis design and airflow CFD analysis</td>
</tr>
<tr>
<td>PUE</td>
<td>Power Usage Effectiveness</td>
<td>Measure of datacenter power efficiency</td>
</tr>
<tr>
<td>IOPS</td>
<td>Disk Input/Output Operations per second</td>
<td>Storage subsystem performance analysis</td>
</tr>
<tr>
<td>DB/RH</td>
<td>Dry Bulb: Air temperature Wet Bulb: Temperature indicating amount of moisture in the air</td>
<td>Determining server operating range and acceptable environmental specs</td>
</tr>
</tbody>
</table>
Microsoft Datacenters: Providing services 24x7

- **Windows Live ID**: More than 1 billion authentications per day
- **bing**: More than 2 billion queries per month
- **Exchange Hosted Services**: Processes 2–4 billion e-mails per day
- **Windows Live Messenger**: 320 million active accounts
- **msn**: 550 million unique visitors monthly
- **Windows Live Hotmail**: 400 million active accounts
Microsoft Datacenters – Global scaling!

"Datacenters have become as vital to the functioning of society as power stations." – The Economist

Quincy, Washington: approx 500K sq ft, 27MW, uses entirely hydro-electric power
San Antonio, Texas: approx 477K sq ft, 27MW, uses recycled water for cooling
Chicago, Illinois: 707,000 square feet with critical power of 60 MW, uses water side economization, containers
Dublin, Ireland: approx 570K sq ft, up to 27MW, uses outside air for cooling.
A large mega-datacenter is 11 times the size of a football field
Design factors for Datacenter Infrastructure

Datacenter and Server architecture
  – Power distribution efficiency
  – Cost efficiency
  – Thermal design analysis

Platform architecture
  – Application performance analysis
  – New platform architecture exploration

Reliability analysis
  – Environmental operating ranges and impact on MTBF

Bringing it all together – Holistic Systems Design
Design factors for Datacenter Infrastructure

Datacenter architecture
- Power distribution efficiency
- Cost efficiency
- Thermal design analysis

Platform architecture
- Application performance analysis
- New platform architecture exploration

Reliability analysis
- Environmental operating ranges and impact on MTBF

Bringing it all together – Holistic Systems Design
For a typical large mega-datacenter...

Building costs are between $10M to $15M per MegaWatt

Typical industry PUE ranges from 1.5-2.0

Energy Consumption: US power rate 10.27 cents per Kilowatt hour) according to DOE/eia

Facility level distribution options

Consider the following common topologies ...

- **UPS Double conversion or Line interactive**
  - AC to AC
  - 480Vac
  - 480/277Vac
  - 480/277Vac
  - 600Vac

- **PDU/ Xfmr**
  - AC to AC
  - 480Vac
  - 415/240Vac
  - 480/277Vac
  - 600Vac

- **PSU**
  - AC to DC
  - 208Vac
  - 240Vac
  - 277Vac
  - 208Vac

- **Server**
  - DC
  - 12Vdc
  - 12Vdc
  - 12Vdc
  - 12Vdc

- **UPS Double conversion or Line interactive**
  - AC to DC
  - 480Vac
  - 480Vac
  - 480/277Vac

- **PDU**
  - DC
  - 575Vdc
  - 380Vdc
  - 48Vdc

- **PSU**
  - DC
  - 48Vdc
  - 380Vdc
  - 48Vdc

- **Server**
  - DC
  - 12Vdc
  - 12Vdc
  - 12Vdc
  - 12Vdc
Analyzing Facility level distribution efficiency

Which topology is the most efficient (lowest PUE)?

Efficiency depends on config type AND Load level
DC configs prevail at low loads, AC configs prevail at high loads
Highest efficiency AC and DC configs are within 1-2% of each other

Source: Green Grid (http://www.thegreengrid.org/~media/WhitePapers/White_Paper_16_-_Quantitative_Efficiency_Analysis_30DEC08.ashx)

Kushagra Vaid, SLAML'10

Oct 3, 2010
Challenges with datacenter power allocation

- Problem statement: Given the above power load curve, how would you maximize server density for the datacenter power envelope
- Tradeoffs:
  - Provision based on peak load → stranded power from variable load traffic patterns
  - Provision based on average load → risk of tripping circuit breakers at higher load levels

Oct 3, 2010

Kushagra Vaid, SLAML'10
Challenges with datacenter power allocation

- Short duration power spikes from sudden load increases occur in reality
- Server heterogeneity and app power profile changes also need to be considered
- Power Capping can be used, but requires understanding tradeoff to performance SLAs and may require sophisticated policy management
- The eventual allocation is a calculated risk to maximize server density without stranding power, taking load patterns into consideration
Design factors for Datacenter Infrastructure

Datacenter architecture
   – Power distribution efficiency
   – **Cost efficiency**
   – Thermal design analysis

Platform architecture
   – Application performance analysis
   – New platform architecture exploration

Reliability analysis
   – Environmental operating ranges and impact on MTBF

Bringing it all together – Holistic Systems Design
Datacenter TCO breakdown

- Datacenter build costs are 17% of TCO
- Energy usage costs are 16% of TCO
- Server capex accounts for largest portion of TCO (61%)
  - Invest in mechanisms to improve work done per watt

Assumptions:
- 10MW facility
- PUE 1.25
- $10/W construction costs
- $0.10c/KWhr power costs
- Server: $2000, 200W
- 3yr server amortization
- 15yr datacenter amortization

Source: James Hamilton (http://mvdirona.com/jrh/TalksAndPapers/PerspectivesDataCenterCostAndPower.xlsx)
Microsoft’s Chicago Data Center

- $500M+ investment
- 1.5 million man-hours-of-labor
- 3000 construction related jobs
- 3400 tons of steel
- 707,000 sq ft
- 190 miles of conduit
- 60 MW Total Critical Power
- 2400 tons of copper
- 7.5 miles of chilled water piping
- 26,000 cubic yards of concrete
Microsoft’s Modular Datacenters

- **Airside economization – outside air cooling!**
- Ultra-efficient water utilization
- Focus on renewable materials
- 30-50% more cost effective
- 1.05 – 1.15 PUE
- ITPAC is the “datacenter-in-a-box”
Design factors for Datacenter Infrastructure

**Datacenter architecture**
- Cost efficiency
- Power distribution efficiency
- Thermal design analysis (CFD modeling)

**Platform architecture**
- Application performance analysis
- New platform architecture exploration

**Reliability analysis**
- Environmental operating ranges and impact on MTBF

**Bringing it all together – Holistic Systems Design**
ITPAC design overview (showing thermal modeling and construction)
IT Pre-Assembled-Components (ITPAC)
Design factors for Datacenter Infrastructure

Datacenter architecture
- Cost efficiency
- Power distribution efficiency
- Thermal design analysis

Platform architecture
- Application performance analysis
- New platform architecture exploration

Reliability analysis
- Environmental operating ranges and impact on MTBF

Bringing it all together – Holistic Systems Design
Bing is computationally intensive and exhibits good instruction level parallelism.

- Most memory operations are loads (as expected) for index reads.
- Pipeline stalls due to cache misses are small (~7%) relative to other stall events.
- Analysis helps understand how to optimize code for maximizing QPS (Queries per Second).
Workload scalability analysis

- How does the multicore trend affect online workloads?
- Identify workload scalability bottlenecks if any
- What is the tradeoff for different frequency offerings in commodity CPUs?

Source: Kozyrakis et al, “Server Engineering Insights for Online Services, IEEE Micro, July 2010
E.g. above shows how in-depth storage trace analysis can be used to understand I/O workload patterns and IOPS rates

HDD power models can then be used to determine optimal power provisioning values – minimizing stranded power and allowing higher density
Design factors for Datacenter Infrastructure

Datacenter architecture
- Cost efficiency
- Power distribution efficiency
- Thermal design analysis

Platform architecture
- Application performance analysis
- New platform architecture exploration

Reliability analysis
- Environmental operating ranges and impact on MTBF

Bringing it all together – Holistic Systems Design
Evaluating mobile CPUs (Atom) for Bing

- Concept: Scale down the Bing platform to use ultra-low power mobile CPUs
  - Advantages: Low cost and power
  - Disadvantages: Performance, Significant tuning
- Results below show how the current systems compare
  - Overall, Xeon systems are currently 2.3x better than Atom on a Perf/W/$ basis
  - Atom system power is high, even though CPU power is low (few watts)
  - However, significant room for improvements as Atom-based servers are optimized in the future
CPU pipeline analysis for Atom

- Atom CPU weak on Branches, FP and caches → 3x worse CPI
- Overall TCO is 2.5x worse on both Perf/$ and Perf/Watt
- Need future Atom CPUs to either provide improved perf or much lower power
- **Source: “Web Search Using Mobile Cores”, ISCA 2010**

<table>
<thead>
<tr>
<th>Xeon</th>
<th></th>
<th></th>
<th>Atom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harpertown 4-core, 2-socket</td>
<td>Diamondville 2-core, 1-socket</td>
<td>Hypothetical 8-core, 2-socket</td>
<td></td>
</tr>
<tr>
<td>Cost ($)</td>
<td>Power (W)</td>
<td>Cost ($)</td>
<td>Power (W)</td>
</tr>
<tr>
<td>Processor</td>
<td>760</td>
<td>125</td>
<td>45</td>
</tr>
<tr>
<td>Motherboard</td>
<td>200</td>
<td>30</td>
<td>80</td>
</tr>
<tr>
<td>Network Interface</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Memory (4GB)</td>
<td>150</td>
<td>8</td>
<td>150</td>
</tr>
<tr>
<td>Storage (HDD)</td>
<td>100</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>Total Server Cost</td>
<td>1210</td>
<td>178</td>
<td>375</td>
</tr>
<tr>
<td>Efficiency ×10⁻³</td>
<td>6.38</td>
<td>43.37</td>
<td>2.67</td>
</tr>
</tbody>
</table>

QPS/$ QPS/W QPS/$ QPS/W QPS/$ QPS/W
Improving Perf/W for SQL workloads

• Seek architectural opportunities for **dramatic** Perf/W improvements
• E.g. shown below for SSD caching solutions implemented in HW RAID controllers and in SW buffer pool management algorithms
• Upto ~3x server consolidation possibility for database workloads

*Khessib et al, “Using Solid State Drives as a Mid-Tier Cache in Enterprise OLTP applications”, TPC Technology Conference on Performance Evaluation and Benchmarking (TPC-TC), Sept 2010*
Design factors for Datacenter Infrastructure

Datacenter architecture
  – Cost efficiency
  – Power distribution efficiency
  – Thermal design analysis

Platform architecture
  – Application performance analysis
  – New platform architecture exploration

Reliability analysis
  – Environmental operating ranges and impact on MTBF

Bringing it all together – Holistic Systems Design
ASHRAE recommended range: 64F-81F Drybulb, max 60% RH
However, most server vendors specify 50F-95F, max 90%RH
Higher temperature operation ➔ Lower fan speeds to cool servers ➔ Improved PUE
However, need to take into account reliability aspects for server components
### Temperature sensitivity analysis

- **HDD case temp distribution** shown for a 35-bay JBOD array (over 3 months) for fixed server inlet temperature (25C)
- **Note increase in ΔT** for inner drive bays
- **Airside economization scenarios** may imply higher inlet temperatures - **What are implications to HDD MTBF?**

**Oct 3, 2010** Kushagra Vaid, SLAML'10

### Table: HDD temperature distribution

| HDD-TEMP | B-1 | B-2 | B-3 | B-4 | B-5 | B-6 | B-7 | B-8 | B-9 | B-10 | B-11 | B-12 | B-13 | B-14 | B-15 | B-16 | B-17 | B-18 | B-19 | B-20 | B-21 | B-22 | B-23 | B-24 | B-25 | B-26 | B-27 | B-28 | B-29 | B-30 | B-31 | B-32 | B-33 | B-34 |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 27 C     | 6.7 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 28 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 29 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 30 C     | 1.3 | 0.3 | 37.4 | 15.1 | 22.9 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 31 C     | 60.0 | 29.4 | 56.0 | 56.7 | 74.3 | 0.1 | 38.4 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 32 C     | 53.3 | 37.4 | 40.6 | 5.8 | 20.3 | 56.3 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 33 C     | 1.4 | 2.9 | 1.3 | 5.1 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 34 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 35 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 36 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 37 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 38 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 39 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 40 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 41 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 42 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 43 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 44 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 45 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 46 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 47 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 48 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 49 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 50 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 51 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 52 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 53 C     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |

- HDD case temp distribution shown for a 35-bay JBOD array (over 3 months) for fixed server inlet temperature (25C)
- **Note increase in ΔT** for inner drive bays
- Airside economization scenarios may imply higher inlet temperatures - **What are implications to HDD MTBF?**
Modeling HDD AFR sensitivity to temperature

• Model the HDD Failure Acceleration Factor, adjusted for production failure data (left figure)

• Create mapping based on temperature distribution data (figure below)

• Calculate overall AFR for a given input server inlet temperature

Oct 3, 2010
Kushagra Vaid, SLAML'10

Failure Acceleration Factor (norm. to 40°C, 100% duty cycle)
Design factors for Datacenter Infrastructure

Datacenter architecture
  – Cost efficiency
  – Power distribution efficiency
  – Thermal design analysis

Platform architecture
  – Application performance analysis
  – New platform architecture exploration

Reliability analysis
  – Environmental operating ranges and impact on MTBF

**Bringing it all together – Holistic Systems Design**
Optimal systems design (Infrastructure perspective)

Multi-dimensional optimization problem

- Perf Requirements
- Power Budget
- Reliability/Availability
- Application architecture
- Cost tradeoffs
- Operational requirements
- Management/Provisioning
- Site-specific environmental
- Various other requirements
Research areas

- Optimal Power provisioning for high dynamic range workloads (idle at night, peak load during day)
- Power aware task scheduling on large clusters
- Energy proportionality via system architecture innovations
- Reliability model for entire datacenter, correlated for parameters such as temperature, humidity, load levels
- Failure prediction opportunities based on log analysis of continuous feeds from management consoles

- Several others...
Summary

• Datacenter design and optimization for various application scenarios involves several disciplines with complex interactions

• An optimal design is usually a delicate balance between various tradeoffs – no easy answers

• Extensive data analysis is the key to making effective design choices

• Several areas for improved design and reliability via data mining and predictive analysis
Q & A