Choosing the Right Technology Platform for the Building of Watson

Patrick O’Rourke
pmorour@us.ibm.com
Executive Briefing Center
Dates: February 14 / 15 / 16 2011

Competition with humans at the game of Jeopardy:

Competition: Two most successful Jeopardy contestants of all time
Watson Overview

Watson History.
- 3+ years development by IBM scientists
- IBM Research Software Stack

Why Jeopardy?
- Grand challenge for a computing system
- Broad range of subject matter,
- Speed at which contestants must provide both accurate responses
- Determine a confidence they are correct
Choosing the Right Technology Platform
Factors Affect Choosing the Right Platform

Would you purchase a family car solely on one factor?

<table>
<thead>
<tr>
<th>Car</th>
<th>Server Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purchase price</td>
<td>Purchase price</td>
</tr>
<tr>
<td>Gas mileage, cost of repairs, insurance cost</td>
<td>Cost of operation, Power consumption, Floor space</td>
</tr>
<tr>
<td>Reliability</td>
<td>Reliability</td>
</tr>
<tr>
<td>Safety, maneuverability, visibility, vendor service</td>
<td>Availability, Disaster recovery, Vendor service</td>
</tr>
<tr>
<td>Storage capacity, number of seats, towing capacity</td>
<td>Scalability &amp; Throughput</td>
</tr>
<tr>
<td>Horsepower</td>
<td>Chip performance</td>
</tr>
<tr>
<td>Dash board layout, Steering wheel location</td>
<td>Instrumentation &amp; Skills</td>
</tr>
<tr>
<td>Handling, comfort, features</td>
<td>Manageability</td>
</tr>
<tr>
<td>Driving Needs / Requirement</td>
<td>People hauler, Carry goods, Usage, Driving Conditions</td>
</tr>
<tr>
<td>Looks, styling, size</td>
<td>Peer and Industry recognition</td>
</tr>
</tbody>
</table>

Would you purchase a family car solely on one factor?
General Workload Attributes … …

Transaction Processing and Database
- High Transaction Rates
- High Quality of Service
- Peak Workloads
- Resiliency and Security

Analytics and High Performance
- Compute or I/O intensive
- High memory bandwidth
- Floating point
- Scale out capable

Business Applications
- Scale
- High Quality of Service
- Large memory footprint
- Responsive infrastructure

Web, Collaboration and Infrastructure
- Highly threaded
- Throughput-oriented
- Scale out capable
- Lower Quality of Service
Main Watson Software Components (Linux Based)

Watson system used **UIMA** as its principal infrastructure for component interoperability.

Made extensive use of the **UIMA-AS** scale-out capabilities that can exploit modern, highly parallel hardware architectures.

UIMA manages all work flow and communication between processes, which are spread across the cluster.

**Apache Hadoop** manages the task of preprocessing Watson's enormous information sources by deploying UIMA pipelines as Hadoop mappers, running UIMA analytics.

**DeepQA**, a “Collection of Algorithms” that can be divided into independent parts, each executed by a separate processor. Computation is embarrassing parallel. It gathers, evaluates, weighs and balances different types of evidence, delivering the answer with the best support it can find.
Watson Apps Environment…

22 Different Process Types
- Heavily Parallelized

389 Processes
- 199 C++
- 190 Java

Threads were heavily core multi-threaded
- Threads had various memory requirements

Watson:
Processes 80 trillion operations (teraflops) per second
Accessing 200 million pages of content
- Against 6 million logic rules to “Understand” the nuances, meanings, and patterns in spoken human language
Data Sources

Sources of information for Watson:

- Encyclopedias
- Dictionaries
- Thesaurus
- Newswire articles
- Literary works
- Bible
- Databases, taxonomies, and ontologies.
  - IMDb
  - DBpedia
  - Wordnet
  - YAGO
- Wikipedia (Full text)

200 million pages of structured and unstructured content

- 1 Million books
- > 4 TB of disk storage
- > Functional memory > 7/8 TB of memory
DeepQA Computing: Memory Intensive

Generates and scores many hypotheses using a combination of 1000’s **Natural Language Processing, Information Retrieval, Machine Learning and Reasoning Algorithms.**

- Question & Topic Analysis
- Question Decomposition
- Hypothesis Generation
- Hypothesis and Evidence Scoring
- Synthesis
- Final Confidence Merging & Ranking

built on **UIMA** for interoperability

built on **UIMA-AS** for scale-out and speed

100s Possible Answers

1000’s of Pieces of Evidence

100,000’s scores from many simultaneous Text Analysis Algorithms

Potential Answers
Workload characterization....

1. **Data Intensive** – Large working set and/or high I/O content applications

2. **I/O Bound** – e.g. High I/O content applications

3. **Mixed Low** – e.g. Multiple, data-intense applications or skewed OLTP, MQ

4. **Mixed High** – e.g. Multiple, cpu-intensive simple applications

5. **Database** – e.g. Oracle DBMS or dynamic HTTP server

6. **Java Light** – e.g. Data intensive java applications

7. **Java Heavy** – e.g. CPU intensive java applications

8. **Skewless OTLP** – e.g. Simple and predictable transaction processing

9. **Protocol Serving** – e.g. Static HTTP, firewall, etc.

10. **CPU Intensive** – e.g. Numerically intensive, etc.
Defining Platform Requirements

- Problem Size
- Time Horizon
- ISV Support
- Nonfunctional Requirements (Availability, Scalability, Security, etc)
- Organization Size
- Flexibility
- Hardware Capabilities
- System x
- Power
- System z
- Hybrid
- Politics
- Skills
- TCO Model
- Strategic Direction and Standards
- Power, cooling, floor space constraints
Choosing the Right Platform for Watson…

System z™
CISC

POWER™
RISC

System x™
x86

Throughput
Quality of Service
Resource Utilization
System Virtualization

Performance
Scalability
Work / Resource / HPC
Resource Virtualization

Standardization
X86 Performance
X86 Scalability
Lowest HW Cost

Scale Up
Scale Out
Choosing the Right Platform?

BlueGene
Blue Gene: Designed to overcome HPC hurdles

- Ultra-scalability for breakthrough science, broad range of HPC applicability
- Parallel workload scaling
- Low power, small footprint, low total cost of ownership (TCO)
- High reliability, 10-100X better MTBF/TF, low maintenance requirements
- Low latency, high performance inter-processor communications
- Reproducible, deterministic runs simplify tracing errors and tuning performance
- High memory bandwidth for data intensive applications
- Open source and standards-based programming environment
Blue Gene continues its leadership performance in a space-saving, power-efficient package for the most performance-demanding applications.

**What is Blue Gene...**

The system scales to 256 racks achieving 3.56 PF/s peak.

**Cores needed:** >24,000
Choosing the Right Platform?

System z
System z

System z™ (CISC)

CPU Performance – workload throughput per resource
Scalability – investment protection
Virtualization – Do more with less
Dynamic – shift resource to workloads
High Resource Utilization - use more of what you own
Reliability – higher service levels
zEnterprise: System of Systems

Managing multi-tier workloads and extending System z governance

**Transaction Processing & Database**
- Application Database
- Data Warehousing
- Online Transaction Processing
- Batch

**Analytics**
- Data Mining Applications
- Numerical
- Enterprise Search

**Business Applications**
- Enterprise Resource Planning
- Customer Relationship Management
- Application Development

**Web, Collaboration and Infrastructure**
- Systems Management
- Web Serving/Hosting
- Networking
- File and Print
IBM zEnterprise System.....

A new dimension in application architecture
Choosing the Right Platform?

Power 775
Power 775 Compute Node

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture</td>
<td>POWER7 256 core per node @ 3.84 GHz</td>
</tr>
<tr>
<td>Cache</td>
<td>On Chip L2 &amp; L3</td>
</tr>
<tr>
<td>DDR3 Memory</td>
<td>Up to 2 TB per node</td>
</tr>
<tr>
<td>PCI Expansion / Node</td>
<td>16 – 16X PCIe Gen 2, 1 - 8X PCIe Gen 2</td>
</tr>
<tr>
<td>Storage Drawers</td>
<td>Up to 6 per rack</td>
</tr>
<tr>
<td></td>
<td>Up to 384 SFF Drives / Drawer</td>
</tr>
<tr>
<td>Ethernet / Node</td>
<td>Up to 16 Quad Port 1 Gb</td>
</tr>
<tr>
<td></td>
<td>Up to 16 Dual Port 10/100</td>
</tr>
<tr>
<td>Remote IO Drawers</td>
<td>None / Internal IO slots</td>
</tr>
<tr>
<td>Cluster Attach</td>
<td>PERCS Fabric</td>
</tr>
<tr>
<td>Power</td>
<td>N+1 Line Cords</td>
</tr>
<tr>
<td>Cooling</td>
<td>Water</td>
</tr>
<tr>
<td>Nodes</td>
<td>Up to 12 per rack</td>
</tr>
</tbody>
</table>
Power 775 – Node Front View

- PCIe Interconnect
- Hub Module (8x)
- Memory DIMMs (64x)
- Water Connection
- 360VDC Input Connector
- P7 QCM (8x)
PERCS POWER7 Hierarchical Structure

- **POWER7 Chip**
  - 8 Cores

- **POWER7 QCM & Hub Chips**
  - QCM: 4 POWER7 Chips
    - 32 Core SMP Image
  - Hub Chip: One per QCM
    - Interconnect QCM, Nodes, and Super Nodes

- **POWER7 IH Node**
  - 2U Node
  - 8 QCMs
    - 256 Cores

- **POWER7 ‘Super Node’**
  - 4 Drawers / Nodes
    - 1024 Cores

- **Full System**
  - Up to 512 ‘Super Nodes’
  - 512K Cores

**Cores needed: 3072**
Choosing the Right Platform?
POWER7 System Highlights

Balance System Design
- Cache, Memory, and IO

POWER7 Processor Technology
- 6th Implementation of multi-core design
- On chip L2 & L3 caches

POWER7 System Architecture
- Blades to High End offerings
- Enhances memory implementation
- PCIe, SAS / SATA, SRIOV

Built in Virtualization
- Cores / Memory and IO
- Mobility Memory Expansion

Workload Flexibility
- Transaction Processing
- ERP Workloads
- High Performance Computing
- Consolidation

Availability
- Processor Instruction Retry
- Alternate Process Recovery
- Hot Add & Services
POWER7 Portfolio

**Major Features:**
- Modular systems with linear scalability
- PowerVM Virtualization
- Physical and Virtual Management
- Roadmap to Continuous Availability
- Binary Compatibility
- Energy / Thermal Management

- **Power 750**
- **Power 720 / 740**
- **Power 710 / 730**
- **Power 770**
- **Power 780**
- **Power 775**
- **Power 755**
- **BladeCenter**
  - PS700 / PS701 / PS702
  - PS703 / PS704
- **AIX 7.1**
- **IBM i 7.1**
## Power 750/755 System

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>POWER7</strong></td>
<td></td>
</tr>
<tr>
<td>POWER7 Architecture</td>
<td>32 Cores @ 3.55 GHz, 88 compute nodes, 2 I/O nodes</td>
</tr>
<tr>
<td>DDR3 Memory</td>
<td>70 nodes 128GB, 30 nodes 256GB</td>
</tr>
<tr>
<td>System Unit SAS SFF Bays</td>
<td>6 HDD 146GB @ 15k</td>
</tr>
<tr>
<td>System Unit IO Expansion Slots</td>
<td>(1) FC1983 dual port 1Gb Ethernet, (1) FC5769 10Gb Fiber SR (RED RIVER) In the 2 I/O nodes: (1) FC5903 SAS RAID Controller, (SQUIB-E)</td>
</tr>
<tr>
<td>Integrated SAS / SATA</td>
<td>Yes</td>
</tr>
<tr>
<td>System Unit Integrated Ports</td>
<td>3 USB, 2 Serial, 2 HMC</td>
</tr>
<tr>
<td>Integrated Virtual Ethernet</td>
<td>Dual 10 Gb HEA</td>
</tr>
<tr>
<td>IO Drawers w/ PCI slots</td>
<td>(4) FC5886 Charlotte SAS enclosure, 2 per I/O node with (12) 300GB cross cabled</td>
</tr>
<tr>
<td>Redundant Power and Cooling</td>
<td>Yes AC Single phase 240 VAC</td>
</tr>
<tr>
<td>EnergyScale</td>
<td>Active Thermal Power Management Dynamic Energy Save &amp; Capping</td>
</tr>
</tbody>
</table>
POWER7 Hardware Selection

Resources

- Needed 2844 cores by June 2010 plus early systems for porting
- 32 core images seemed to match the Jeopardy apps best
- Looked at the following:
  - P7 770: Not dense enough to fit in 10 racks
  - P7 775: Would look the best but schedule did not show it would be stable enough before GA
  - P7 750: Best fit for density
# Jeopardy! System Comparison

## Current P7 750 vs. Potential P7 775 and Blue Gene

<table>
<thead>
<tr>
<th>System Details</th>
<th>Power 750</th>
<th>Power 775</th>
<th>BG/P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frames</td>
<td>10</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Compute Nodes</td>
<td>88</td>
<td>11 Drawers (88 Virtual nodes)</td>
<td>6144 (Each 4 way)</td>
</tr>
<tr>
<td>IO Nodes</td>
<td>2</td>
<td>0.25 Virtual node</td>
<td>48</td>
</tr>
<tr>
<td>Total Cores</td>
<td>2880</td>
<td>3072</td>
<td>24576</td>
</tr>
<tr>
<td>CPU speed</td>
<td>3.55 GHz</td>
<td>~3.8 GHz</td>
<td>850 MHz</td>
</tr>
<tr>
<td>Interconnect</td>
<td>10 Gb Ethernet</td>
<td>HFI</td>
<td>Proprietary</td>
</tr>
<tr>
<td>Est. Total Power</td>
<td>145kW</td>
<td>226kW</td>
<td>240kW</td>
</tr>
<tr>
<td>Cooling Type</td>
<td>Air</td>
<td>Water</td>
<td>Air Optional water</td>
</tr>
<tr>
<td>Frame size</td>
<td>19”</td>
<td>30”</td>
<td>48&quot;x38&quot;</td>
</tr>
</tbody>
</table>
Software

Low Level:
- SLES 11, JAVA, CNFS, GPFS, xCat,

Middleware:
- Apache UIMA (open source)

Applications:
- DeepQA - Main analytical engine which ran on Power 7
- Avatar - Ran on Mac notebook

Voice:
- Synthesis, strategies for betting, buzzing in, clue selection and exchanging info with Jeopardy Computers all ran on Windows 7 Lenovo desktop
Resiliency

Extra nodes added into the system

Two Management Nodes for redundant xCAT cluster management

Dual HMC systems

Parts locker on site for most commonly failed parts

Extra Ethernet switches and blades on hand for final taping

I/O solution was built in redundancy and GPFS software to manage

10 extra systems on site to be used for “Parts”

• Part of these systems were installed as a 6 node test cluster
  • Used to stage new code before applied to the full system
  • Used as a debug platform for bugs in the main system
System Storage

Selection

- **Options: Fiber Channel and SAS Direct Attach**
  - **Fiber Channel**
    - Performance better than SAS (4Gb/s vs 3Gb/s)
    - Good reliability but more parts to fail (controller, disk drawers, switch)
    - Took up more space than SAS
  - **SAS Storage Drawer**
    - Performance was adequate
    - Simple design with direct attach
    - Good RAS with cross cabling between 2 drawers and 2 I/O servers

Hardware

- **I/O Servers**
  - Two Power 750s for redundancy
  - Each 750 has 2 Exp12S drawers with (12) 300GB HDDs
  - “X” cabled between the 2 I/O servers so that each server could access the others disk if disk drawers went down or a server went down
  - Managed by CNFS and GPFS software
Network

Selection Process
- **Cluster network selection**
  - DDR InfiniBand
    - Performance was much more than needed
    - Ethernet had rock solid stability where performance was not a factor
  - 1Gb Ethernet and 10Gb blocking
    - Not enough performance
    - Looked at from a cost perspective
  - Performance was met with non-blocking 10Gb switch

Hardware
- **FSP Network** :
  - (3) 48 port Juniper E48 1Gb Ethernet switch (one not used)
- **Cluster network**
  - (1) Juniper IBM Ethernet Switch J08E (4274-E08)
- **Node adapter**
  - 10Gb HEA was used to interconnect the nodes to the 10Gb switch
  - 10Gb PCIe adapter was installed as a back up
Why Power for Watson/DeepQA?

Workload Optimized System Design
- All components of the stack tuned for throughput and latency
- Tight integration of application specific acceleration technologies (future)

POWER7 Technology
- High single thread performance and throughput capacity
- Large L3 with low latency
- High bandwidth to memory and large shared memory footprint
- High speed SMP fabric and scaling
- High level of reliability and application availability

Watson Stack
- Rapid application development and prototyping
- Open source infrastructure (e.g. UIMA, Hadoop)
- Thousands of parallel processes
Choosing the Right Technology Platform

- Problem Size
- Time Horizon
- ISV Support
- Nonfunctional Requirements (Availability, Scalability, Security, etc)
- Power, cooling, floor space constraints
- Strategic Direction and Standards
- Organization Size
- Flexibility
- Hardware Capabilities
- Politics
- Skills
- TCO Model
Factors Affect Choosing the Right Platform

<table>
<thead>
<tr>
<th>Choosing the Right Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purchase price</td>
</tr>
<tr>
<td>Gas mileage &amp; cost of repairs</td>
</tr>
<tr>
<td>Reliability</td>
</tr>
<tr>
<td>Vendor support services</td>
</tr>
<tr>
<td>Capacity &amp; Options</td>
</tr>
<tr>
<td>Horsepower</td>
</tr>
<tr>
<td>Computing Requirements</td>
</tr>
<tr>
<td>Handling &amp; Features</td>
</tr>
<tr>
<td>Looks, styling, size</td>
</tr>
</tbody>
</table>
The End....