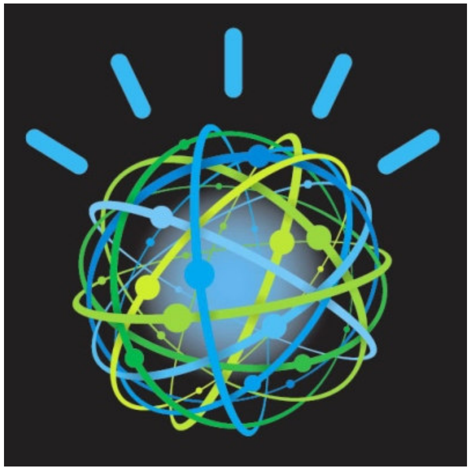


Choosing the Right Technology Platform for the Building of Watson

Patrick O'Rourke
pmorour@us.ibm.com
Executive Briefing Center

Watson takes on Jeopardy!
Advanced computing system has potential to take business intelligence to a new level
JEOPARDY!

- Dates: February 14 / 15 / 16 2011
- Competition with humans at the game of Jeopardy:
  - Human vs. Machine contest.
- Competition: Two most successful Jeopardy contestants of all time

Watson Overview

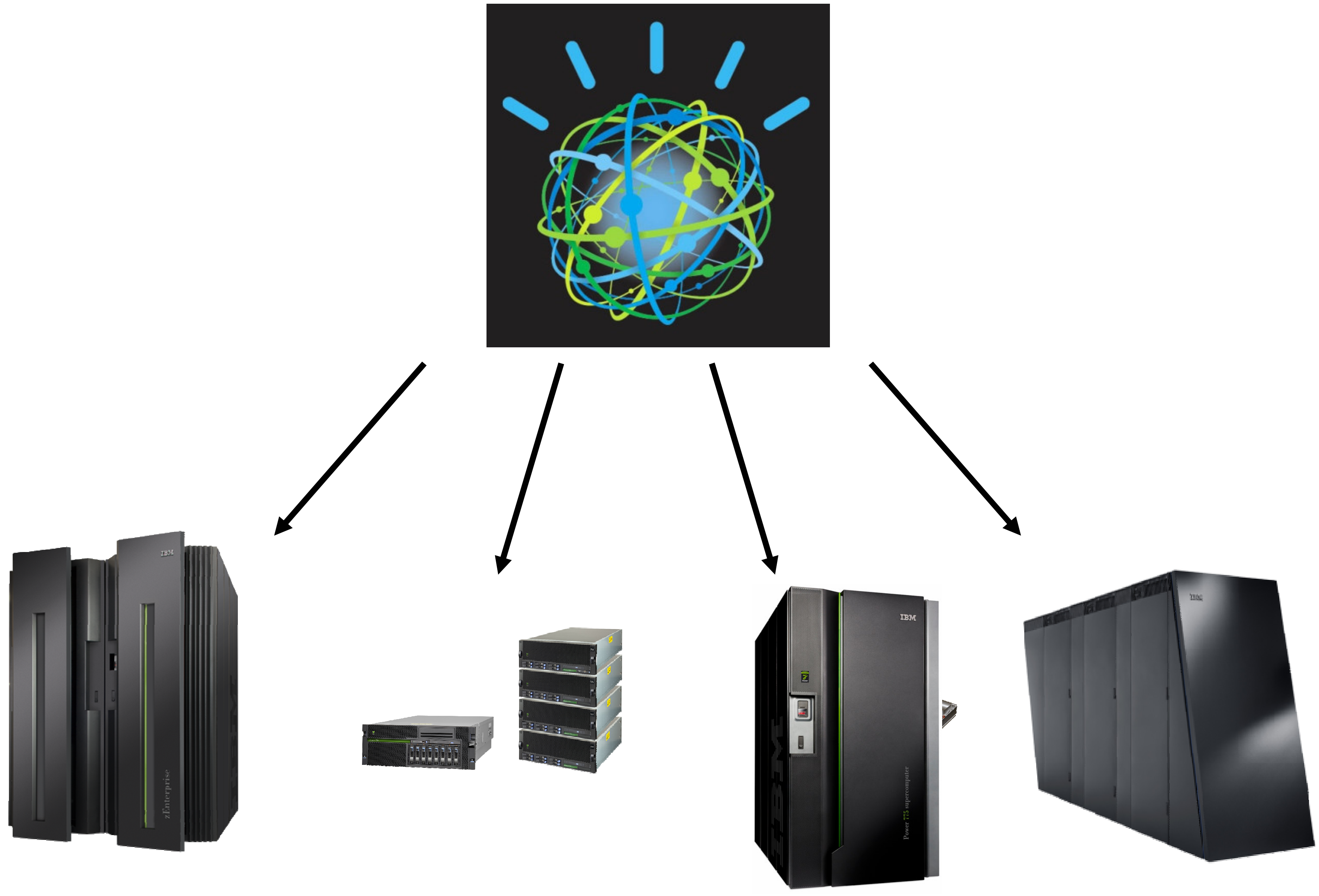
Watson History.

- 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?

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

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:

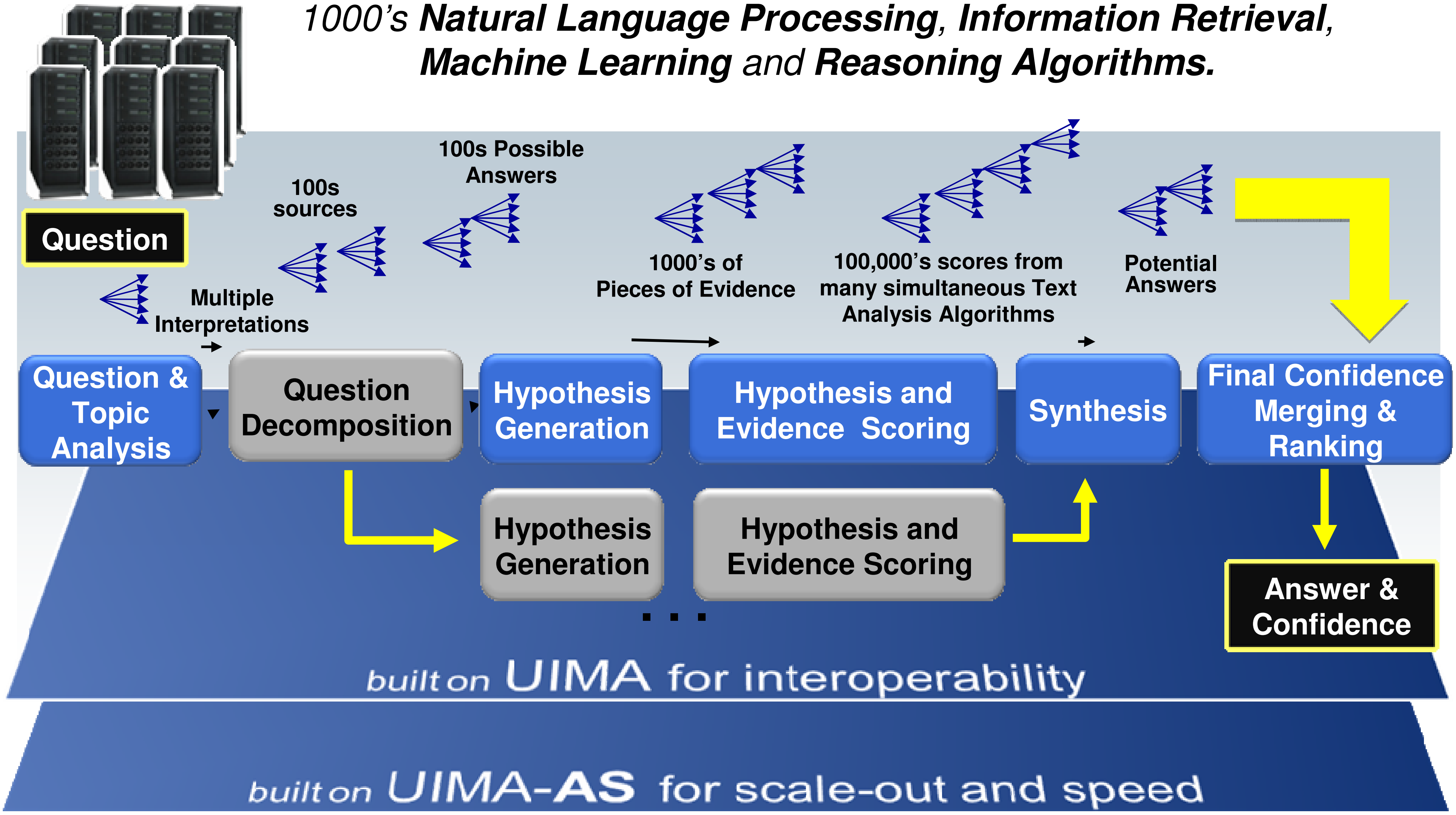
- Encyclopedias
- Dictionaries
- Thesaurus
- Newswire articles
- Literary works
- Bible
- Databases, taxonomies, and ontologies.
  - IMDb
  - DBpedia
  - Wordnet
  - YAGO
- Wikipedia (Full text)

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

Multiple Interpretations

100s sources

100s Possible Answers

1000's of Pieces of Evidence

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

Potential Answers

Question & Topic Analysis

Question Decomposition

Hypothesis Generation

Hypothesis and Evidence Scoring

Synthesis

Final Confidence Merging & Ranking

Hypothesis Generation

Hypothesis and Evidence Scoring

. . .

Answer & Confidence

built on UIMA for interoperability

built on UIMA-AS for scale-out and speed

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-intense simple applications

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

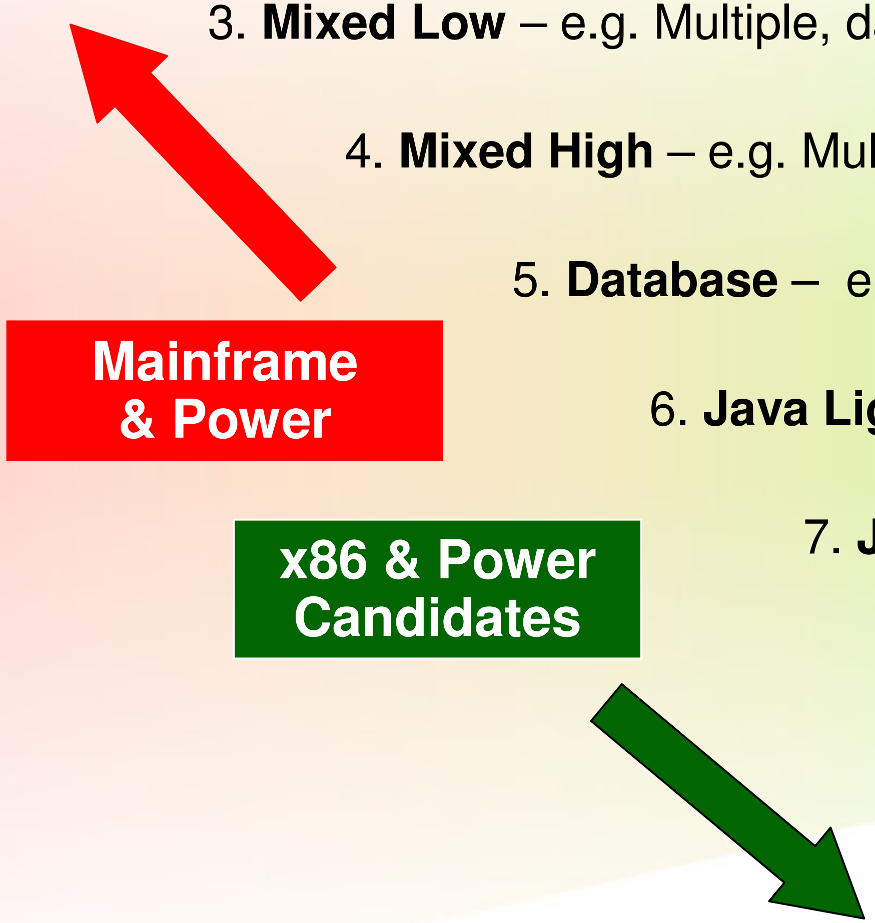
6. **Java Light** – e.g. Data intensive java applications

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

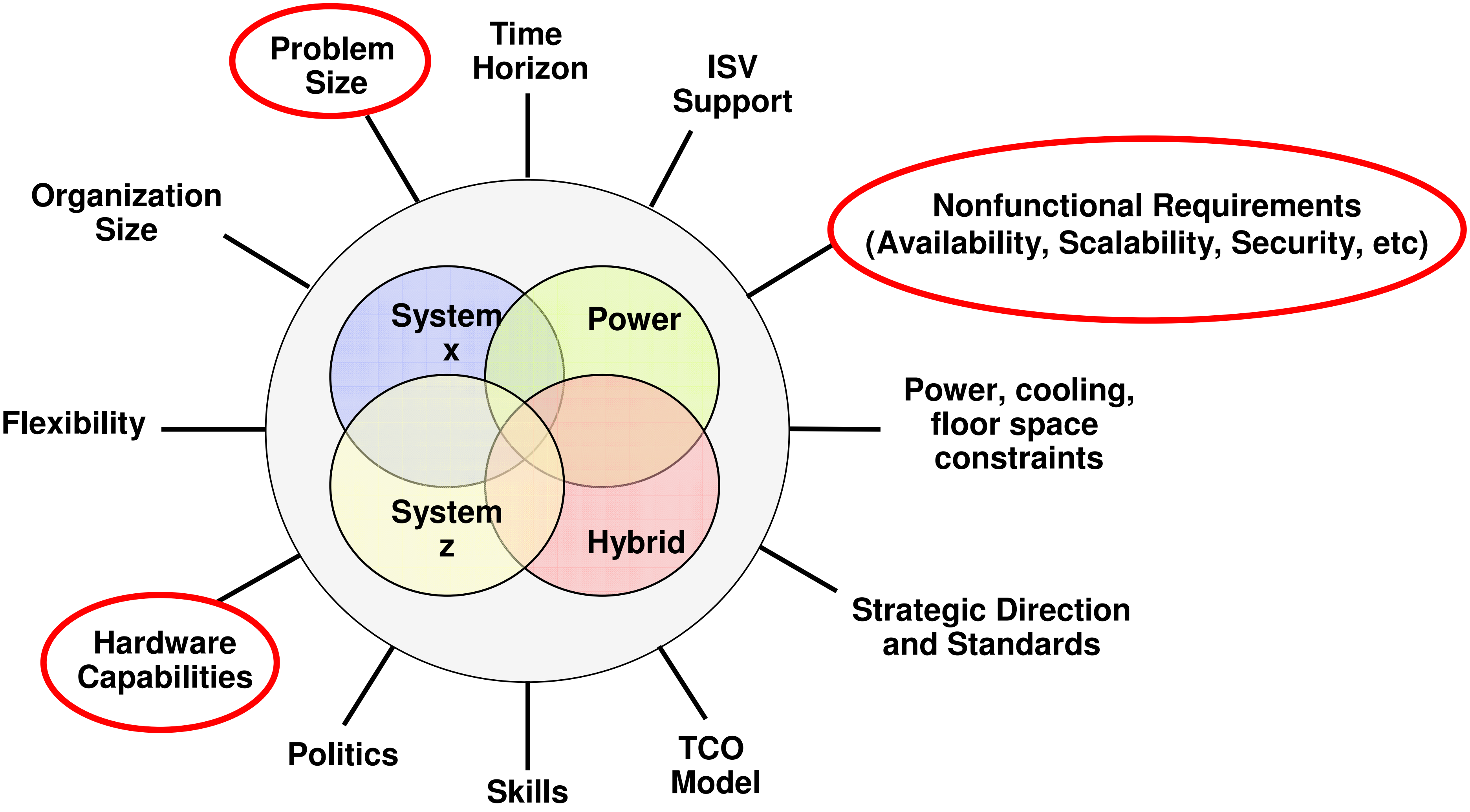
8. **Skewless OTLP** – e.g. Simple and predictable transaction processing

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
Organization Size
Nonfunctional Requirements (Availability, Scalability, Security, etc)
System x
Power
Flexibility
Power, cooling, floor space constraints
System z
Hybrid
Hardware Capabilities
Strategic Direction and Standards
Politics
Skills
TCO Model

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?

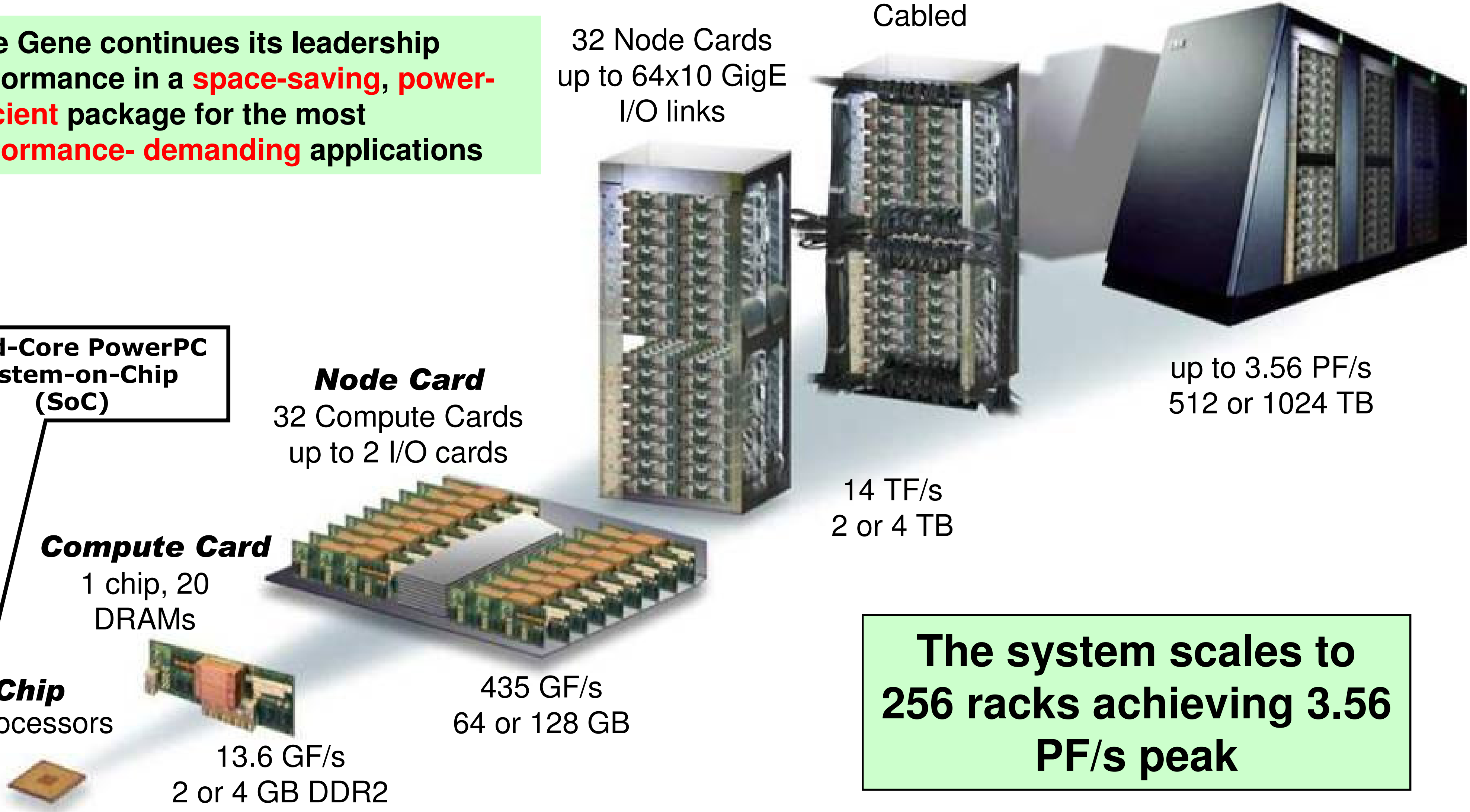
BlueGene

Blue Gene: Designed to overcome HPC hurdles

- Ultra-scalability for breakthrough science, broad range of HPC applicability
- <u>Parallel workload scaling</u>
- Low power, small footprint, low total cost of ownership (TCO)
- High reliability, 10-100X better MTBF/TF, low maintenance requirements
- <u>Low latency, high performance inter-processor communications</u>
- Reproducible, deterministic runs simplify tracing errors and tuning performance
- <u>High memory bandwidth for data intensive applications</u>
- <u>Open source and standards-based programming environment</u>

What is Blue Gene…..

Blue Gene continues its leadership performance in a space-saving, power-efficient package for the most performance- demanding applications



Quad-Core PowerPC System-on-Chip (SoC)

Chip
4 processors

13.6 GF/s
8 MB EDRAM

850 MHz

Compute Card
1 chip, 20 DRAMs

13.6 GF/s
2 or 4 GB DDR2

Node Card
32 Compute Cards
up to 2 I/O cards

435 GF/s
64 or 128 GB

Rack
32 Node Cards
up to 64x10 GigE
I/O links

Cabled

14 TF/s
2 or 4 TB

System
up to 256 racks

up to 3.56 PF/s
512 or 1024 TB

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

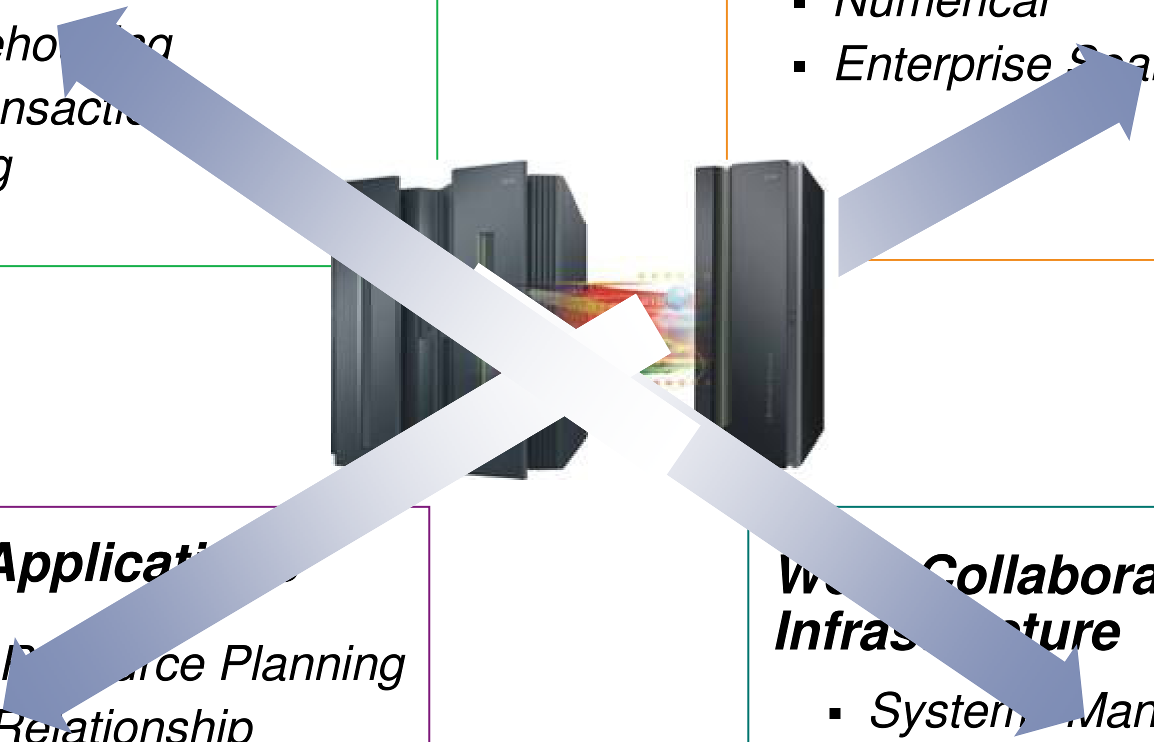
Dynamic – shift resource to workloads

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

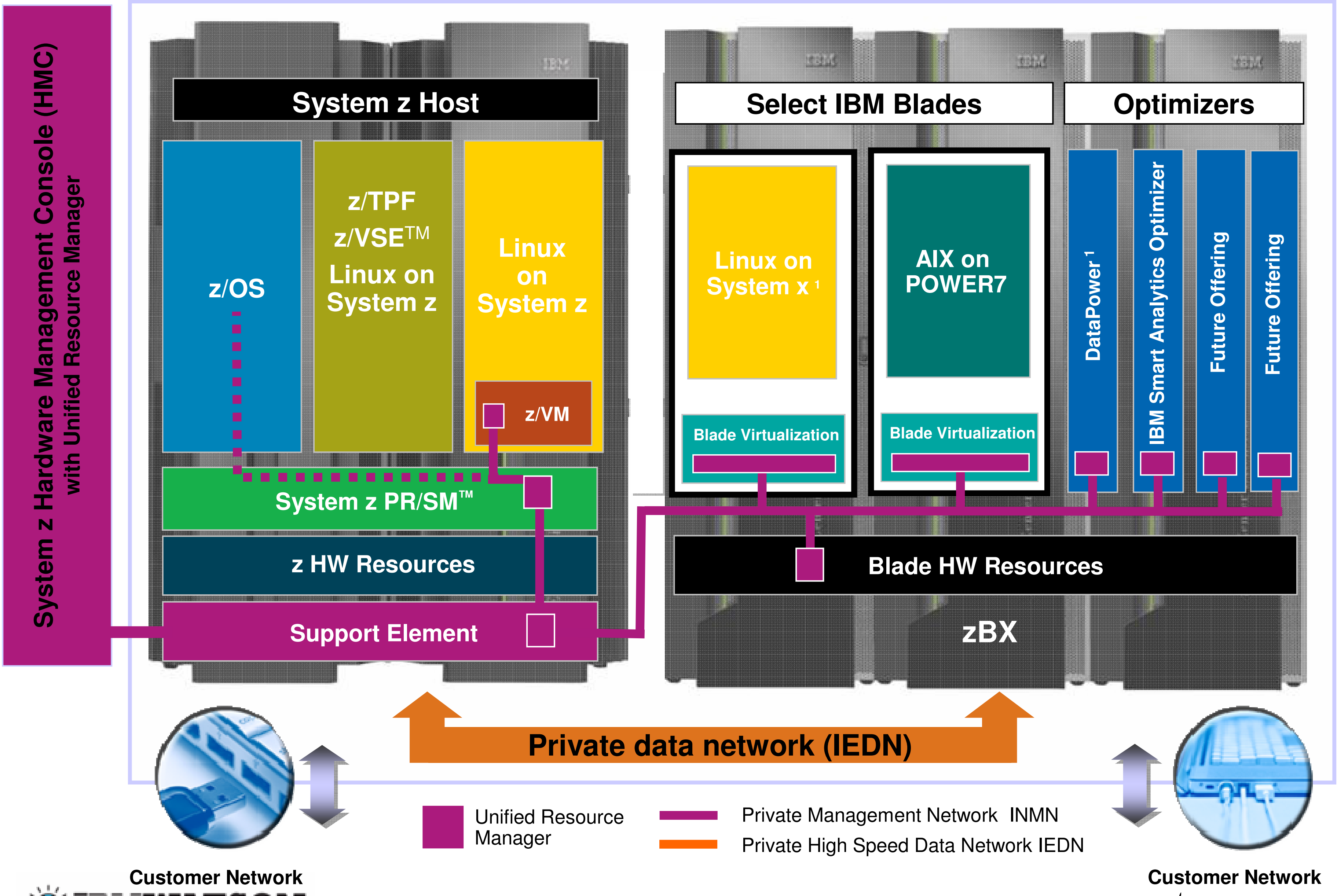
- *Enterprise Resource Planning*
- *Customer Relationship Management*
- *Application Development*

Web, Collaboration and Infrastructure

- *Systems Management*
- *Web Serving/Hosting*
- *Networking*
- *File and Print*

IBM zEnterprise System…..

A new dimension in application architecture



IBM WATSON

Choosing the Right Platform?

Power
775

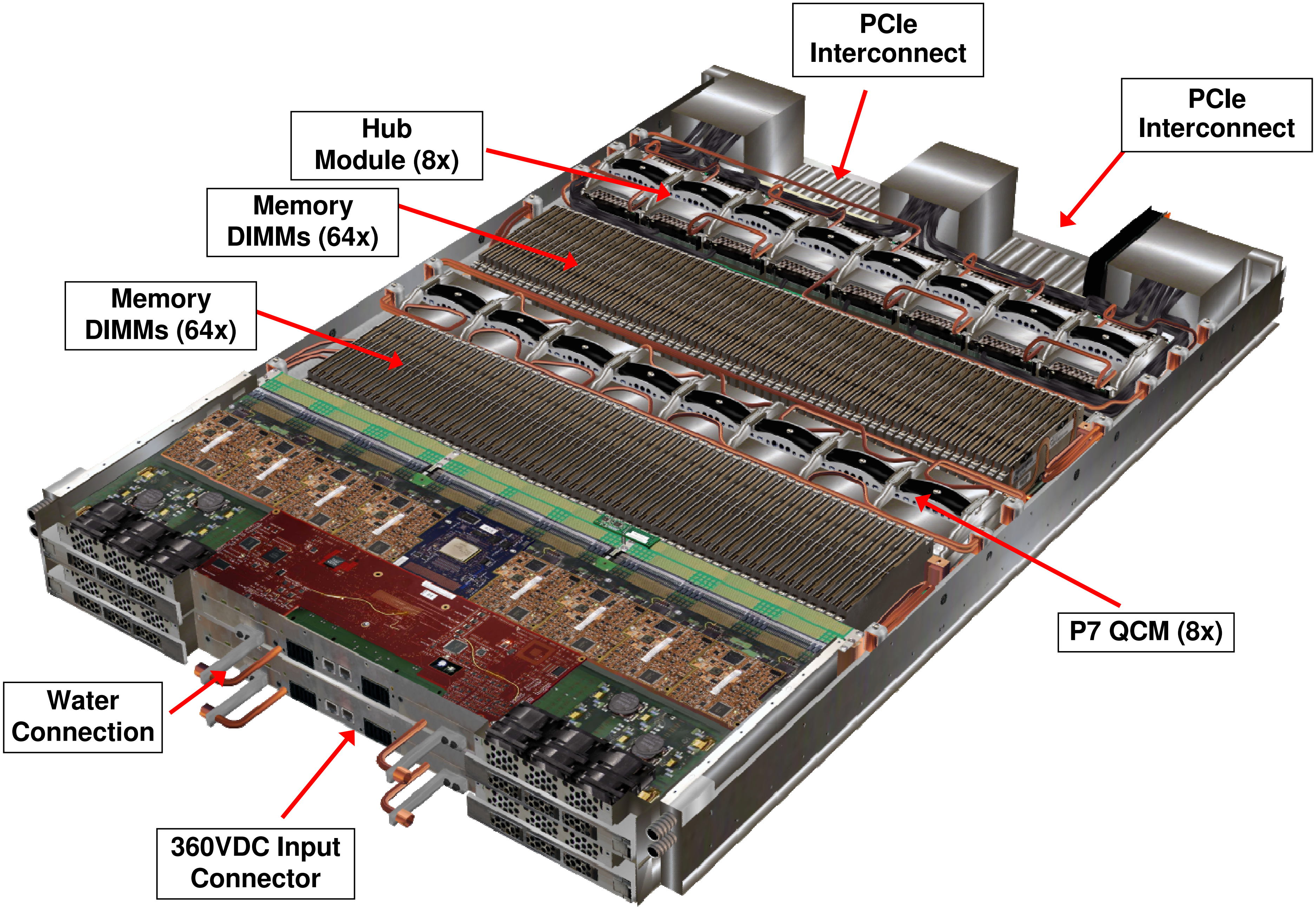
Power 775 Compute Node

AIX

Linux

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

Power 775 – Node Front View



PCIe Interconnect

PCIe Interconnect

Hub Module (8x)

Memory DIMMs (64x)

Memory DIMMs (64x)

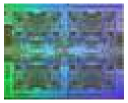
P7 QCM (8x)

Water Connection

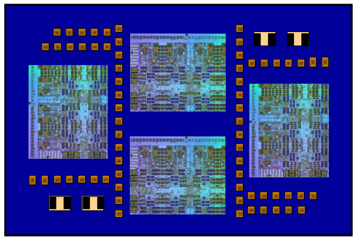
360VDC Input Connector

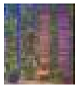
PERCS POWER7 Hierarchical Structure

- **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

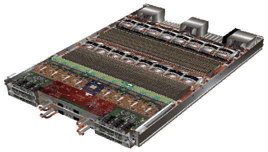




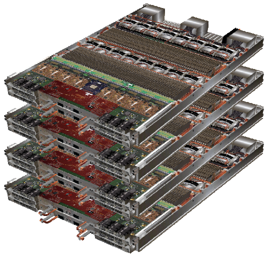
Hub Chip

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

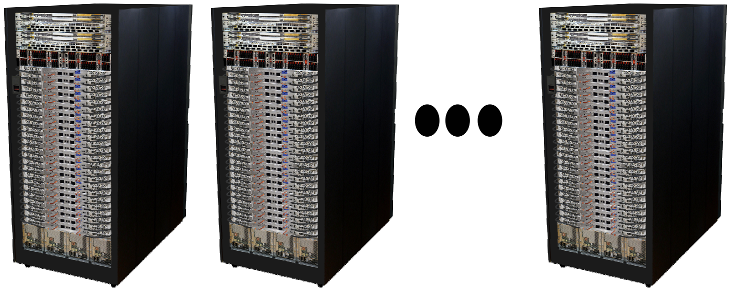
Cores needed: 3072



- **POWER7 'Super Node'**
  - 4 Drawers / Nodes
    - 1024 Cores



- **Full System**
  - Up to 512 'Super Nodes'
  - 512K Cores



Choosing the Right Platform?

Power

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

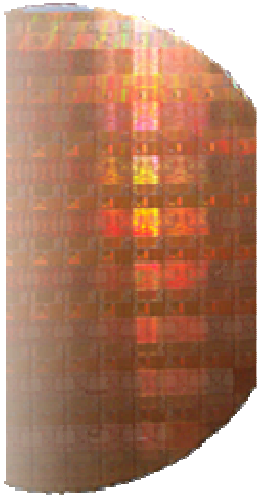
- Cores / Memory and IO
- Mobility Memory Expansion

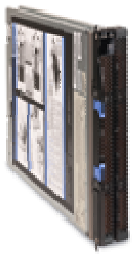
Workload Flexibility

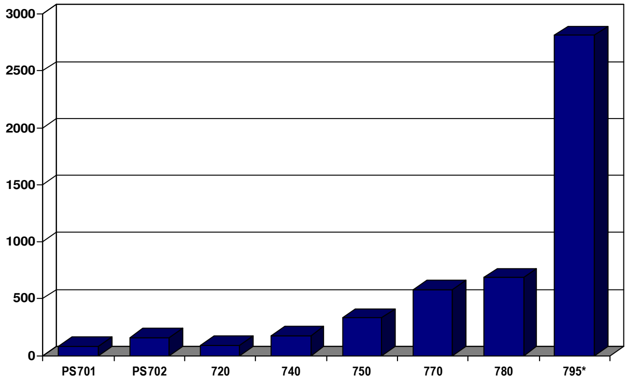
- Transaction Processing
- ERP Workloads
- High Performance Computing
- Consolidation

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 795

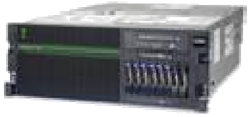
Power 780

Power 770

Power 750

POWER7 BUILT ON Power

Power 720 / 740



Power 710 / 730

BladeCenter
PS700 / PS701 / PS702
PS703 / PS704

PowerHA™ PowerVM™ IBM® Systems Director

Power 775

Power 755

AIX
for Business
Linux

AIX 7.1 IBM i 7.1

Power 750/755 System

POWER7	
POWER7 Architecture	32 Cores @ 3.55 GHz 88 compute nodes 2 I/O nodes
DDR3 Memory	70 nodes 128GB, 30 nodes 256GB
System Unit SAS SFF Bays	6 HDD 146GB @ 15k
System Unit IO Expansion Slots	(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)
Integrated SAS / SATA	Yes
System Unit Integrated Ports	3 USB, 2 Serial, 2 HMC
Integrated Virtual Ethernet	Dual 10 Gb HEA
IO Drawers w/ PCI slots	(4) FC5886 Charlotte SAS enclosure, 2 per I/O node with (12) 300GB cross cabled
Redundant Power and Cooling	Yes AC Single phase 240 VAC
EnergyScale	Active Thermal Power Management Dynamic Energy Save & Capping

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

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

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

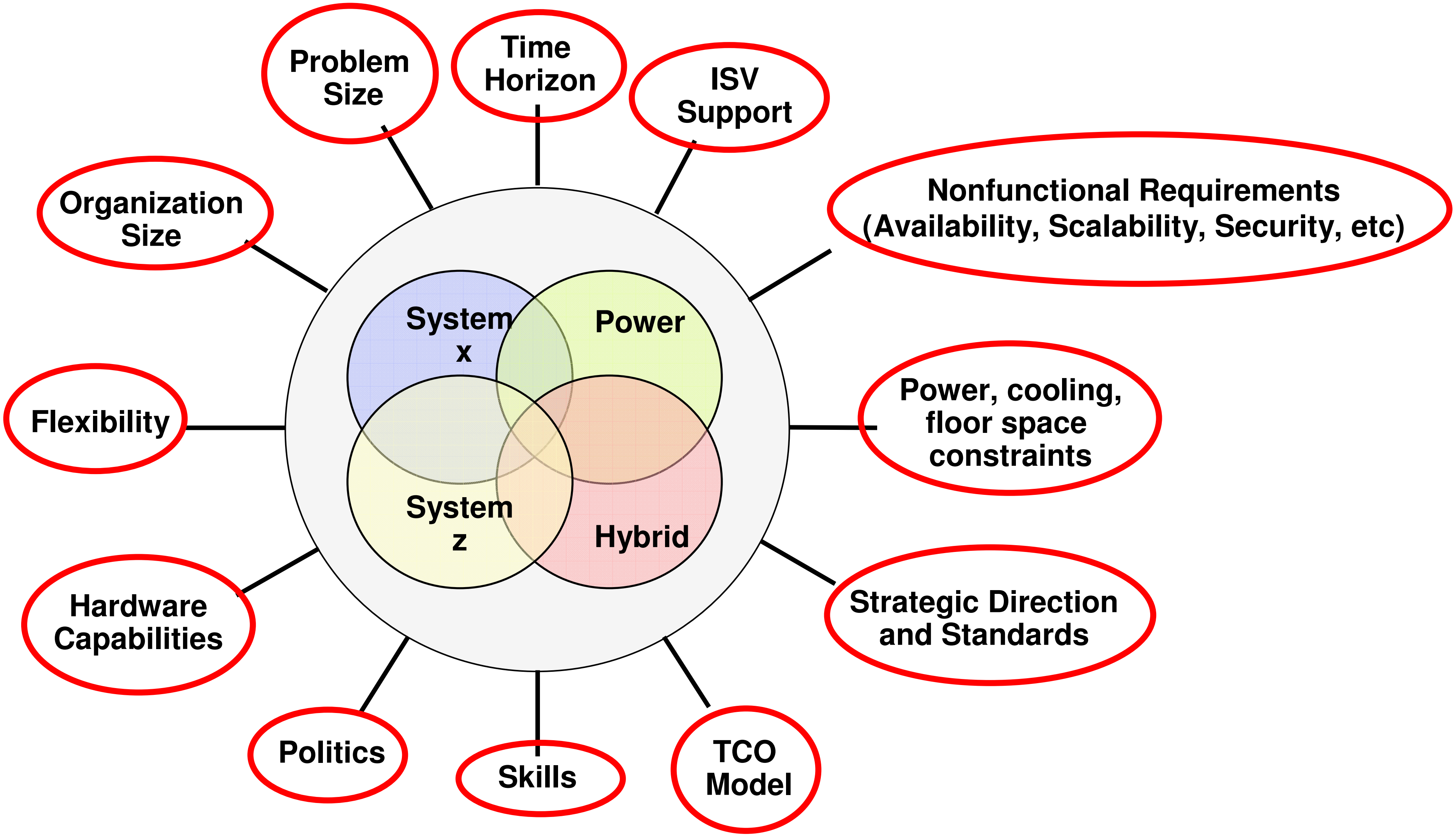
- High single thread performance and throughput capacity
- Large L3 with low latency
- High bandwidth to memory and large shared memory footprint
- High speed SMP fabric and scaling
- High level of reliability and application availability

Watson Stack

- Rapid application development and prototyping
- Open source infrastructure (e.g. UIMA, Hadoop)
- Thousands of parallel processes

Power 750

Choosing the Right Technology Platform .....


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

Factors Affect Choosing the Right Platform

Choosing the Right Platform
Purchase price
Gas mileage & cost of repairs
Reliability
Vendor support services
Capacity & Options
Horsepower
Computing Requirements
Handling & Features
Looks, styling, size

The End….