Ben Newton
Low Latency Technical
Sales

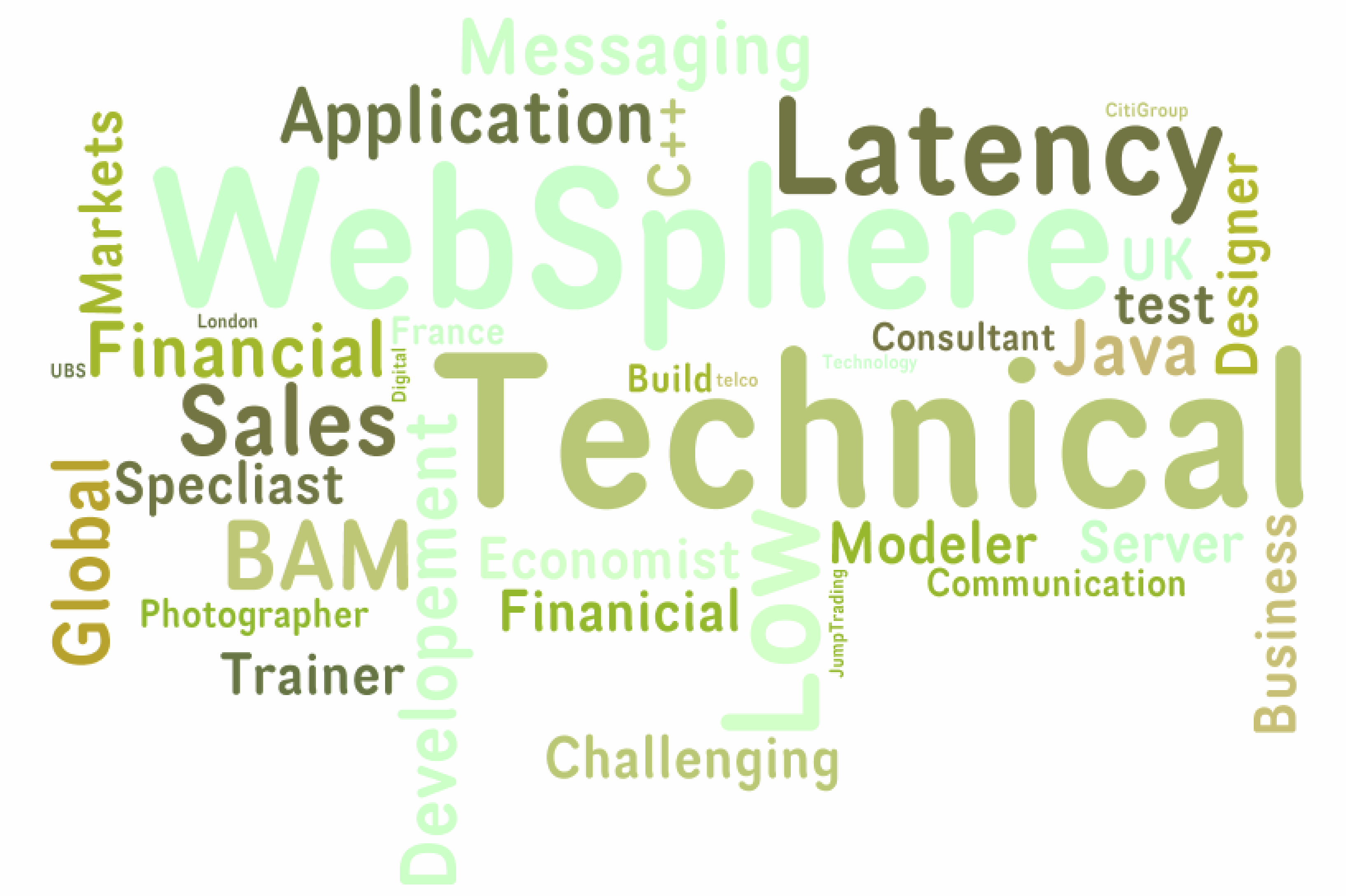
HARDER
STRONGER
FASTER

HARDER > STRONGER > FASTER

Introduction

Life is a drag, but every now and again man gets a chance of greatness

An overview of Ben Newton



Agenda

What is it and who wants it

Lie, benchmarks and statistics

Hybrid Systems

The fastest case study around

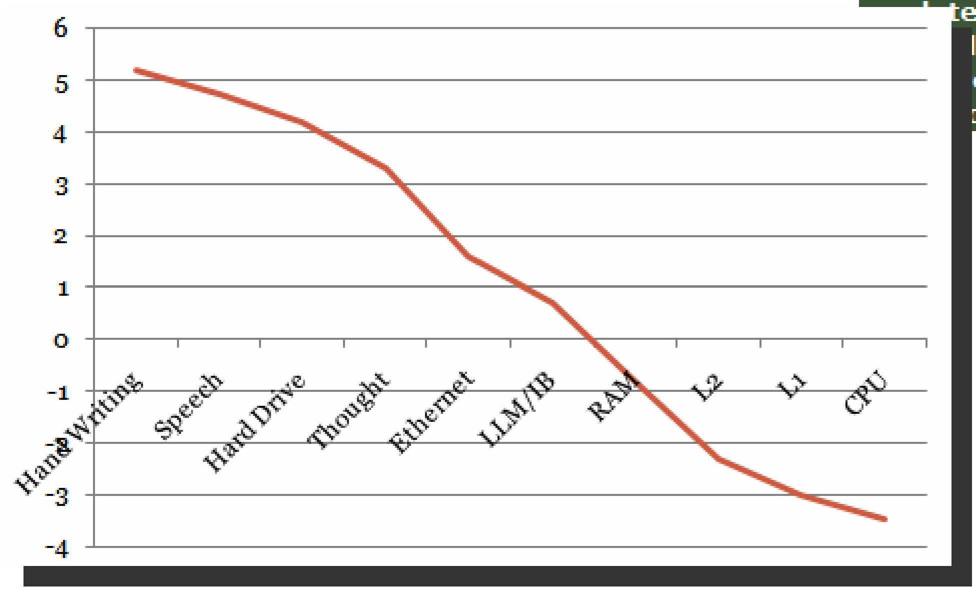
Take Away tips

What are you
looking at?

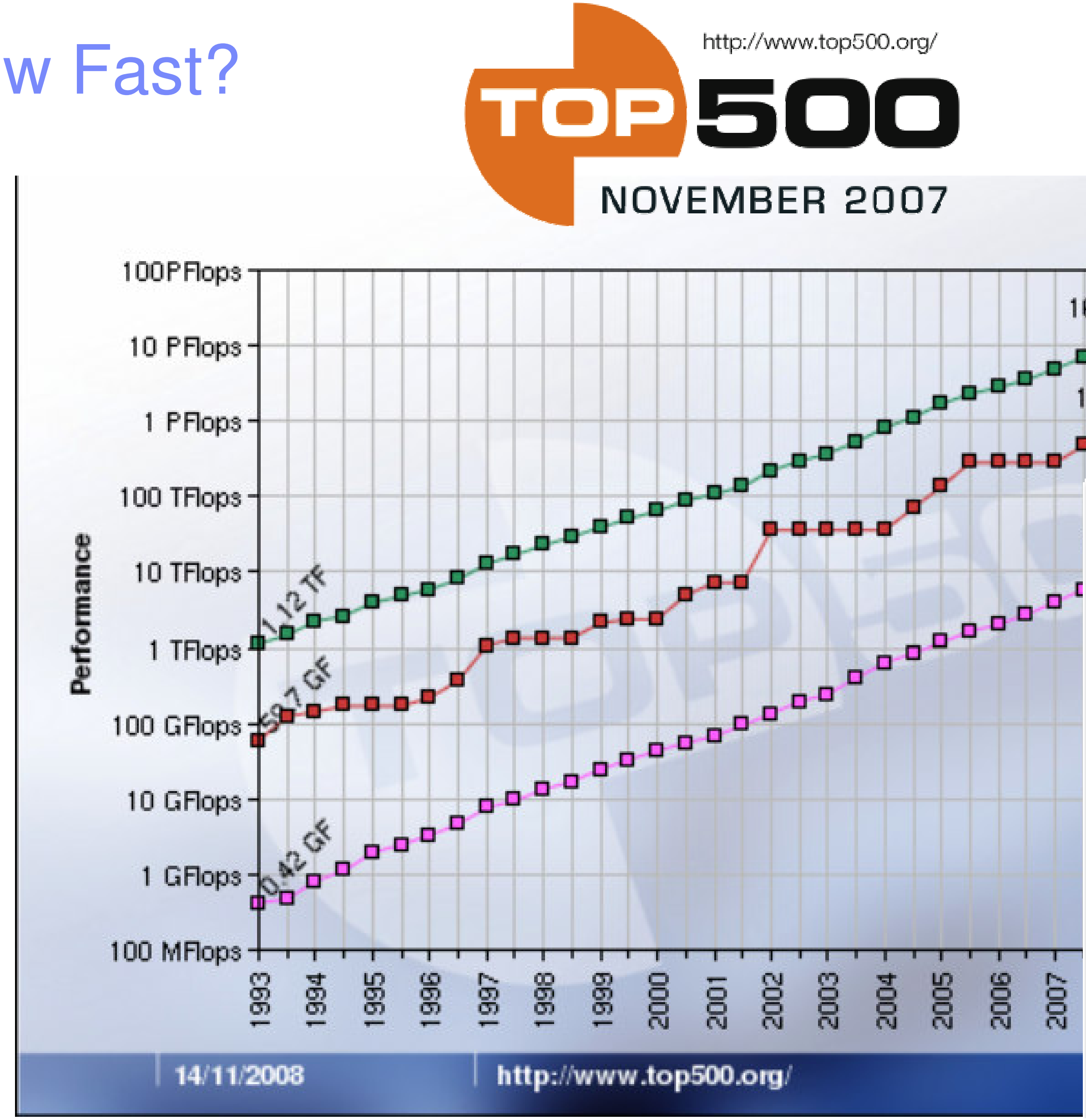
How Fast?

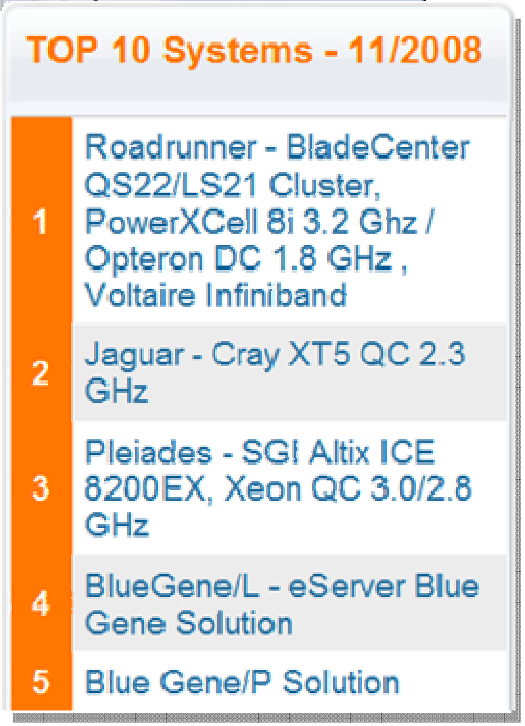
	hand writing	Speech	Hard Drive	Thought	Ethernet	LLM/IB
bytes/s	33	106	60,000,000	500	30,000,000	3,000,000,000
latency	0.15s	50ms	15ms	10ms	38µs	5µs
latency equalled in micro (µ) seconds	150000	50000	15000	10000	38	5

	RAM	L2	L1	CPU
bytes/s	100,000,000			
latency	0.16µs	4.7ns	1ns	0.33ns
tency lled in o (µ) onds	0.16	0.0047	0.001	0.00033



How Fast?



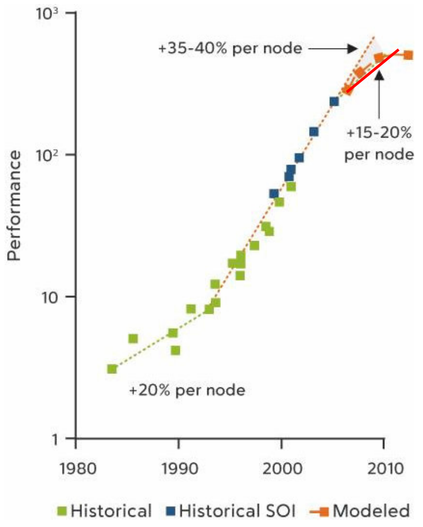


TOP 10 Systems - 11/2008

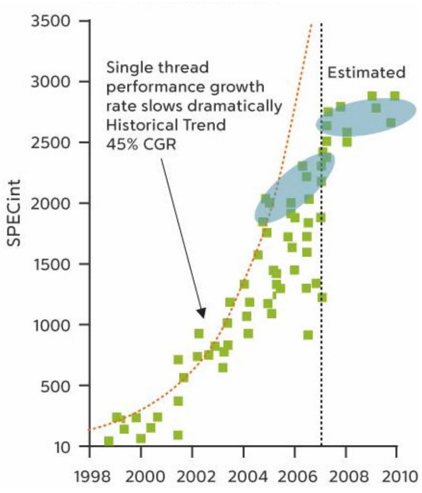
1	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz , Voltaire Infiniband
2	Jaguar - Cray XT5 QC 2.3 GHz
3	Pleiades - SGI Altix ICE 8200EX, Xeon QC 3.0/2.8 GHz
4	BlueGene/L - eServer Blue Gene Solution
5	Blue Gene/P Solution

Compute-power becomes abundant

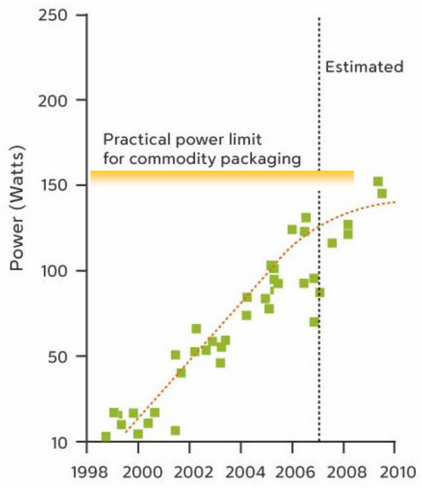
Transistor performance scaling continues, but at a slower rate



Single thread performance is slowing dramatically



Power is limiting practical performance



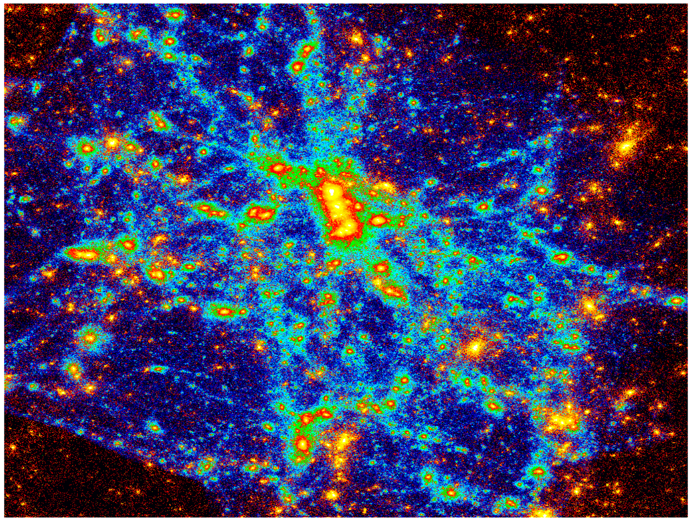
HARDER > STRONGER > FASTER

How reliable, how costly?

Power Consumption & Heat Generation Hurt:
  Reliability, Availability, & Total Cost of Ownership

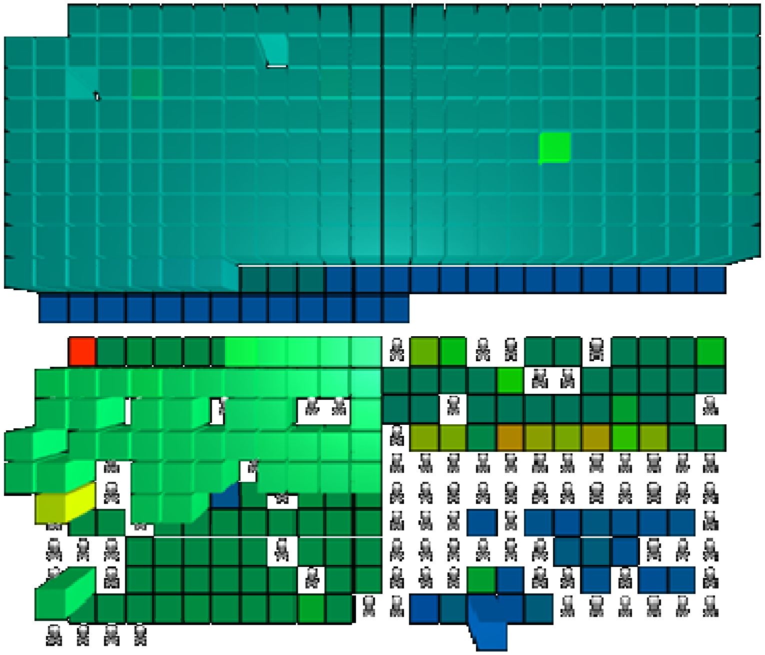
Electrical Power for Computing Costs Money
  Earth Simulator:
    12 MW/year → $10M/year
  World's Processors:
    13 GW/year → $10B/year

N-Body Gravitational Computation

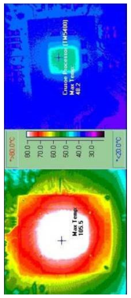


Green Destiny a TOP500

240-Node Supercomputer in 5 Sq. Ft. in 3.2kW Power



Green CPU Array

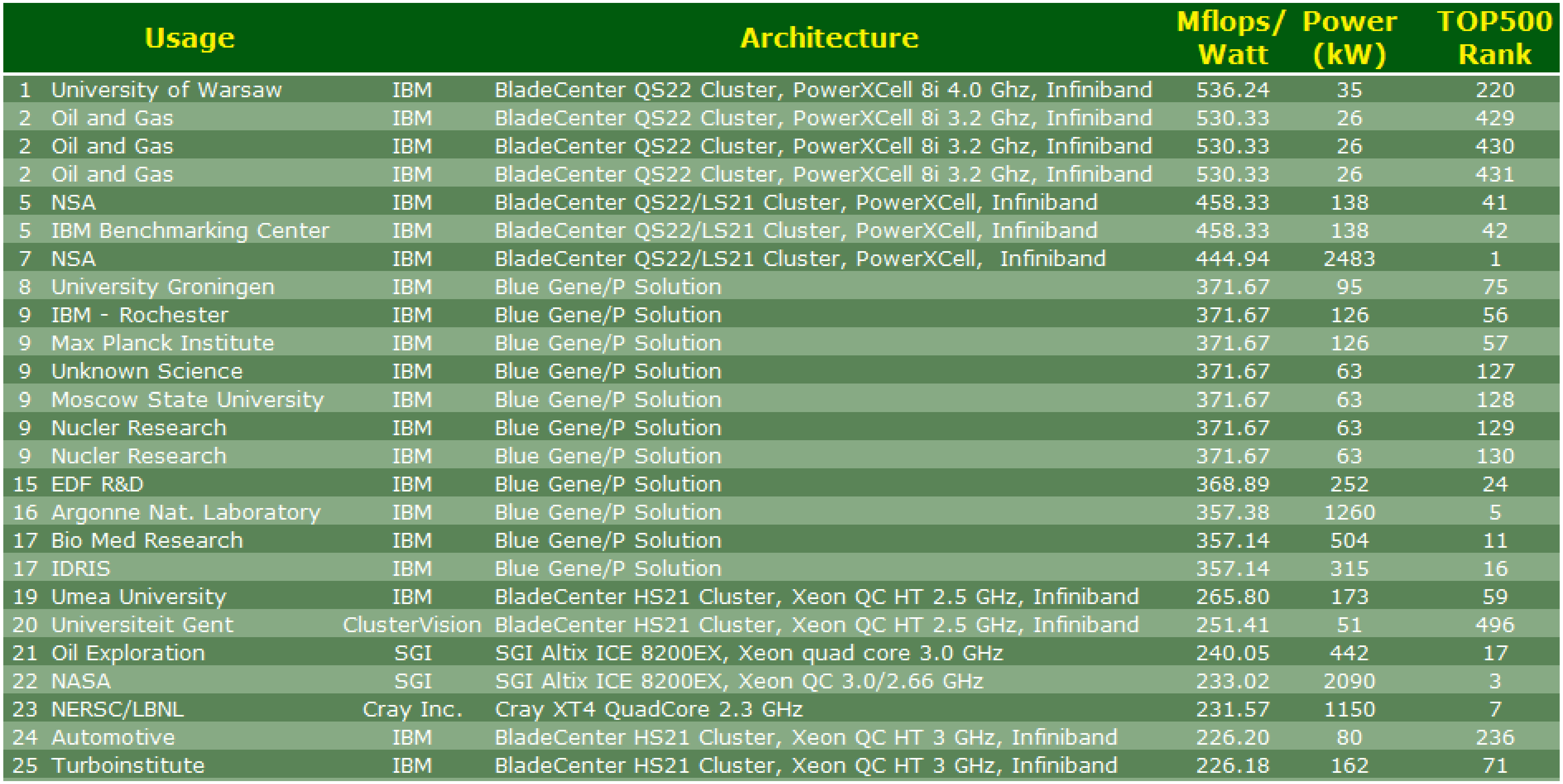


Always On CPU's

Reliability

- ✓ Operating Environment: *A dusty 25°-30°C warehouse.*
- ✓ *No unscheduled downtime in its 24-month lifetime.*

But performance / watt is growing 40% per year



	Usage		Architecture	Mflops/ Watt	Power (kW)	TOP500 Rank
1	University of Warsaw	IBM	BladeCenter QS22 Cluster, PowerXCell 8i 4.0 Ghz, Infiniband	536.24	35	220
2	Oil and Gas	IBM	BladeCenter QS22 Cluster, PowerXCell 8i 3.2 Ghz, Infiniband	530.33	26	429
2	Oil and Gas	IBM	BladeCenter QS22 Cluster, PowerXCell 8i 3.2 Ghz, Infiniband	530.33	26	430
2	Oil and Gas	IBM	BladeCenter QS22 Cluster, PowerXCell 8i 3.2 Ghz, Infiniband	530.33	26	431
5	NSA	IBM	BladeCenter QS22/LS21 Cluster, PowerXCell, Infiniband	458.33	138	41
5	IBM Benchmarking Center	IBM	BladeCenter QS22/LS21 Cluster, PowerXCell, Infiniband	458.33	138	42
7	NSA	IBM	BladeCenter QS22/LS21 Cluster, PowerXCell, Infiniband	444.94	2483	1
8	University Groningen	IBM	Blue Gene/P Solution	371.67	95	75
9	IBM - Rochester	IBM	Blue Gene/P Solution	371.67	126	56
9	Max Planck Institute	IBM	Blue Gene/P Solution	371.67	126	57
9	Unknown Science	IBM	Blue Gene/P Solution	371.67	63	127
9	Moscow State University	IBM	Blue Gene/P Solution	371.67	63	128
9	Nucler Research	IBM	Blue Gene/P Solution	371.67	63	129
9	Nucler Research	IBM	Blue Gene/P Solution	371.67	63	130
15	EDF R&D	IBM	Blue Gene/P Solution	368.89	252	24
16	Argonne Nat. Laboratory	IBM	Blue Gene/P Solution	357.38	1260	5
17	Bio Med Research	IBM	Blue Gene/P Solution	357.14	504	11
17	IDRIS	IBM	Blue Gene/P Solution	357.14	315	16
19	Umea University	IBM	BladeCenter HS21 Cluster, Xeon QC HT 2.5 GHz, Infiniband	265.80	173	59
20	Universiteit Gent	ClusterVision	BladeCenter HS21 Cluster, Xeon QC HT 2.5 GHz, Infiniband	251.41	51	496
21	Oil Exploration	SGI	SGI Altix ICE 8200EX, Xeon quad core 3.0 GHz	240.05	442	17
22	NASA	SGI	SGI Altix ICE 8200EX, Xeon QC 3.0/2.66 GHz	233.02	2090	3
23	NERSC/LBNL	Cray Inc.	Cray XT4 QuadCore 2.3 GHz	231.57	1150	7
24	Automotive	IBM	BladeCenter HS21 Cluster, Xeon QC HT 3 GHz, Infiniband	226.20	80	236
25	Turboinstitute	IBM	BladeCenter HS21 Cluster, Xeon QC HT 3 GHz, Infiniband	226.18	162	71

How do I do it?

Hybridise

Clustered, specialised hardware and software

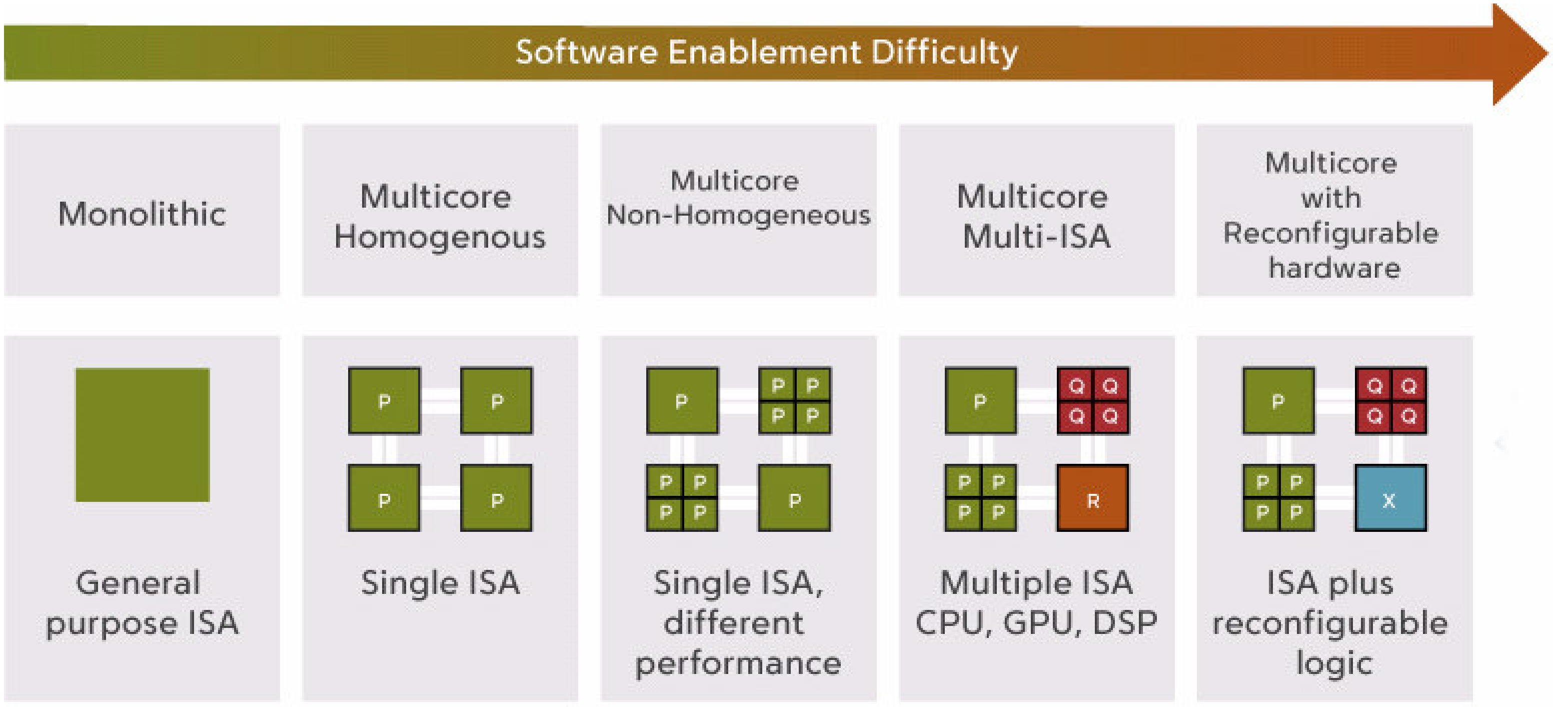
Componentise

Flash Memory
MQ Low Latency Messaging

Optimise

The initial SPEC benchmark addresses only one subset of server workloads: the performance of server side Java.

Software Enablement Difficulty Scale

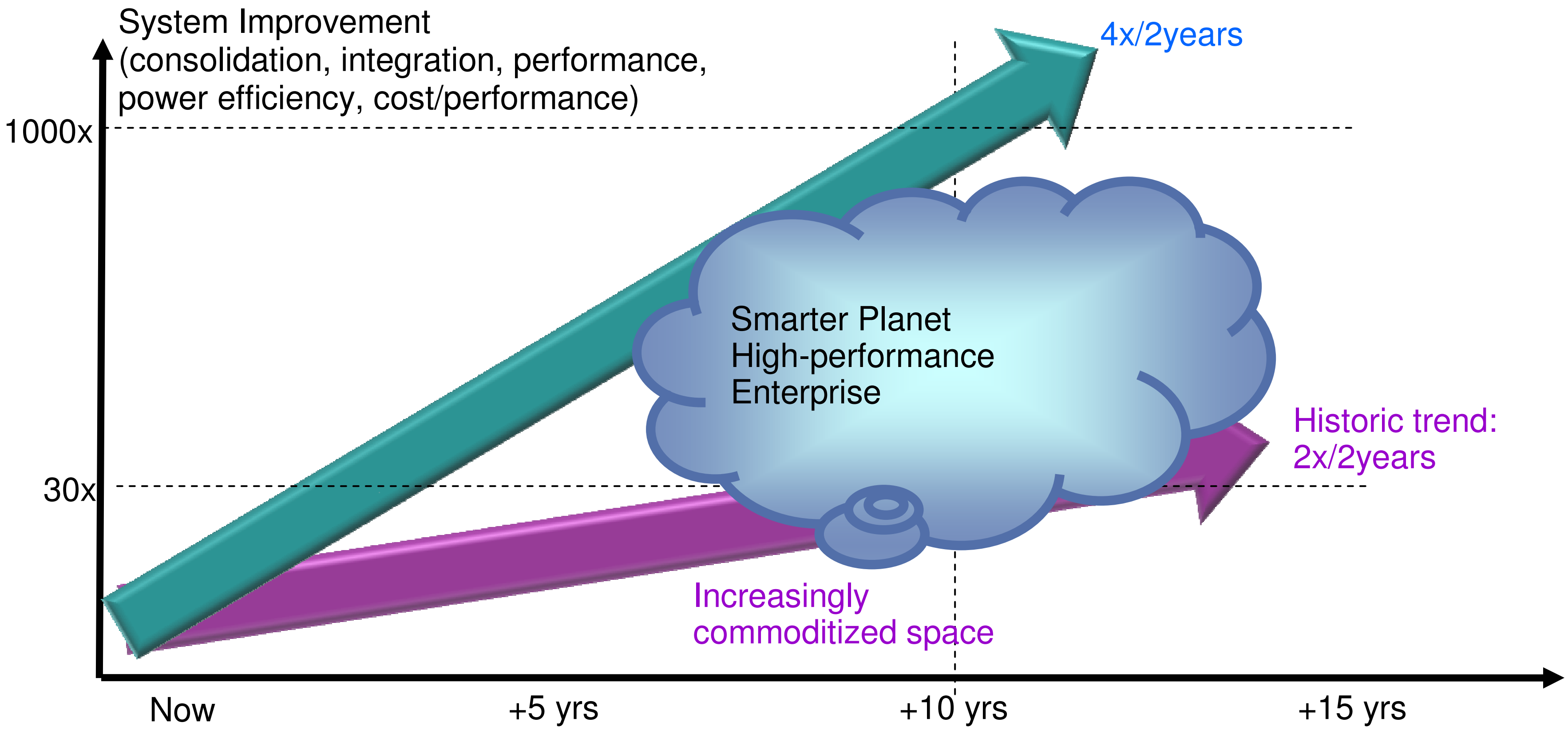


Software Enablement Difficulty

Monolithic	Multicore Homogenous	Multicore Non-Homogeneous	Multicore Multi-ISA	Multicore with Reconfigurable hardware
General purpose ISA	Single ISA	Single ISA, different performance	Multiple ISA CPU, GPU, DSP	ISA plus reconfigurable logic

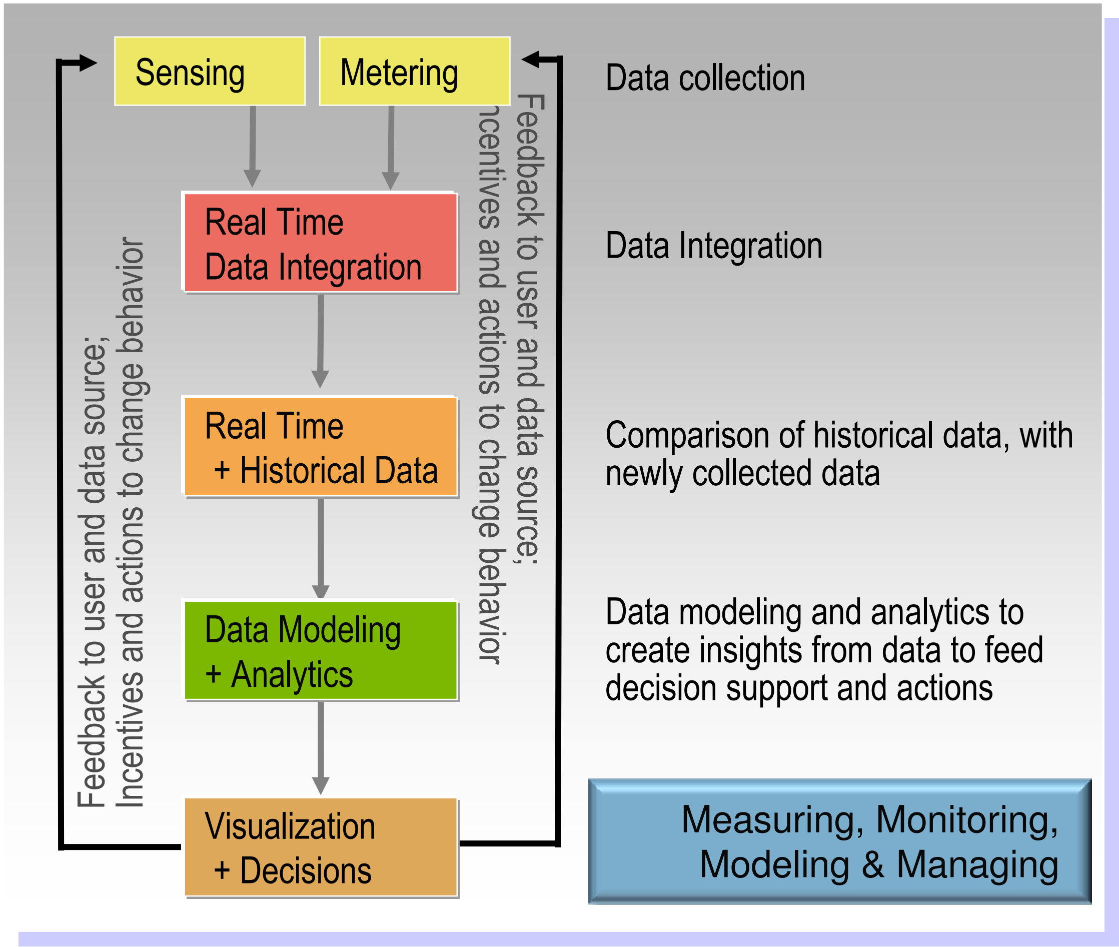
Hybrid Systems

A large class of emerging applications (Smarter Planet, high-performance enterprise), for which network-speed processing and data/compute intensive modeling and simulation are an integral component, will require significant improvement in systems characteristics (consolidation, integration, performance, power efficiency, cost/performance). These applications represent a significant growth opportunity.



Transformational Hybrid Systems

Smarter Planet represents a new paradigm. It applies to multiple business situations, relying on mathematics and models to drive the business activity (for example traffic management, intelligent utility network, etc.). These applications represent a significant opportunity outside the space addressed by conventional commercial system capabilities.



Sensing

Metering

Real Time Data Integration

Real Time + Historical Data

Data Modeling + Analytics

Visualization + Decisions

Feedback to user and data source; Incentives and actions to change behavior

Feedback to user and data source; ncentives and actions to change behavior

Data collection

Data Integration

Comparison of historical data, with newly collected data

Data modeling and analytics to create insights from data to feed decision support and actions

Measuring, Monitoring, Modeling & Managing

Traditional Computing

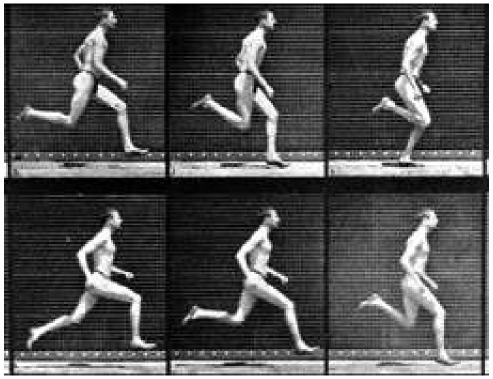
Historical fact finding from data-at-rest

Batch paradigm, pull model

Query-driven: queries against stored data

Relies on Databases, Data Warehouses

Stream Computing



Real time analysis of data-in-motion

Streaming data

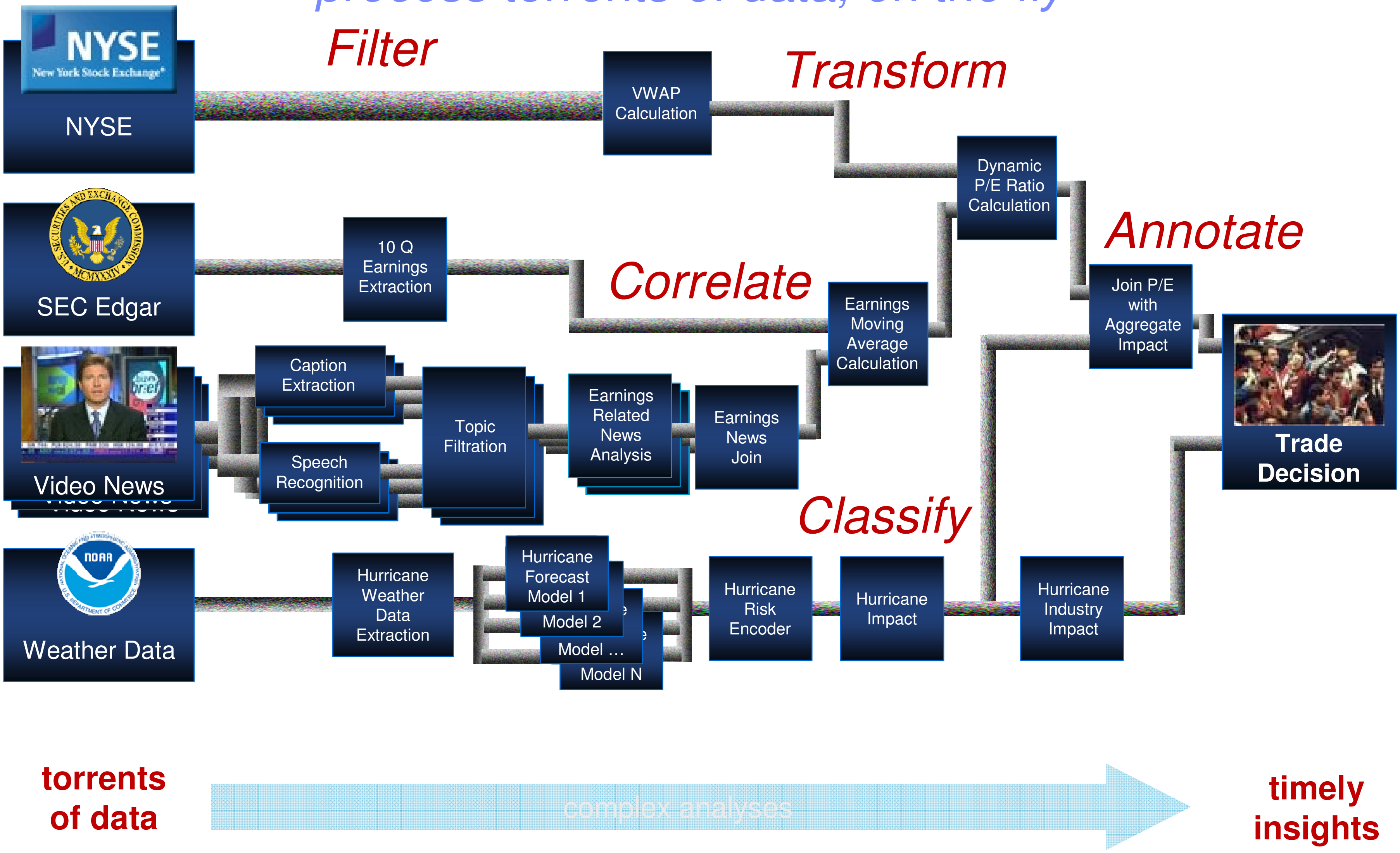
A stream of structured or unstructured data-in-motion

Stream Computing

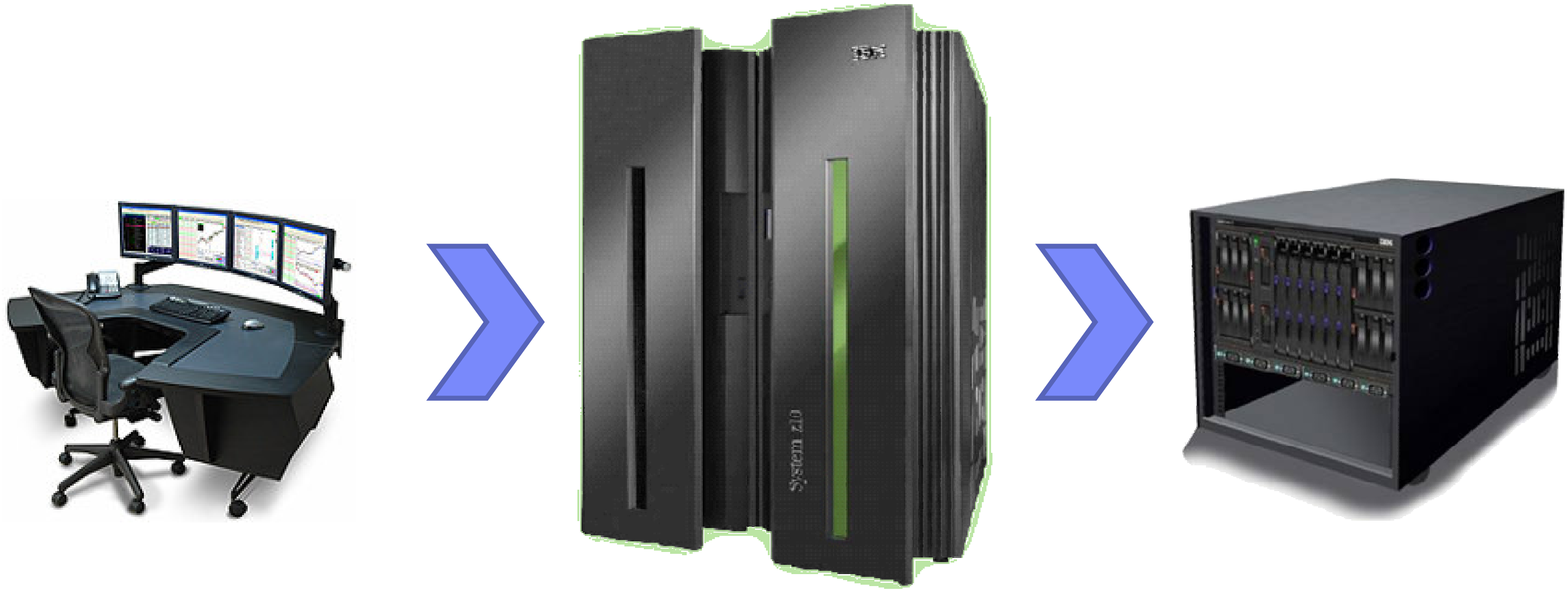
Analytic operations on streaming data in real-time

Scalable Stream Computing

process torrents of data, on the fly



Hybrid Accelerated Analytics

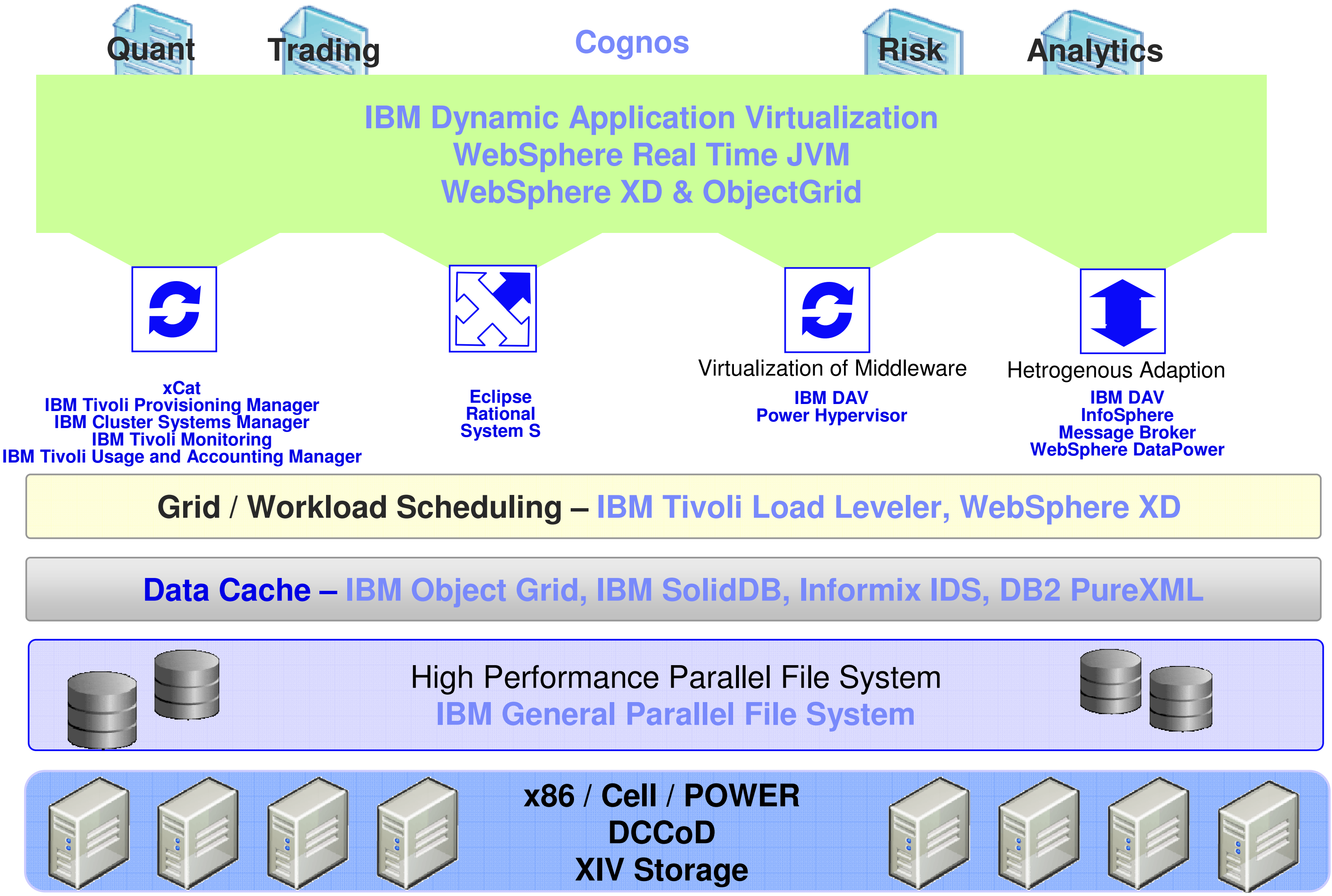


4 x 4core 5GHz

10 x 9cell blades

1.4billion real time option calculations per second

The IBM Hybrid Optimised Analytic Infrastructure



Quant Trading Cognos Risk Analytics

IBM Dynamic Application Virtualization
WebSphere Real Time JVM
WebSphere XD & ObjectGrid

xCat
IBM Tivoli Provisioning Manager
IBM Cluster Systems Manager
IBM Tivoli Monitoring
IBM Tivoli Usage and Accounting Manager

Eclipse
Rational
System S

Virtualization of Middleware
IBM DAV
Power Hypervisor

Hetrogenous Adaption
IBM DAV
InfoSphere
Message Broker
WebSphere DataPower

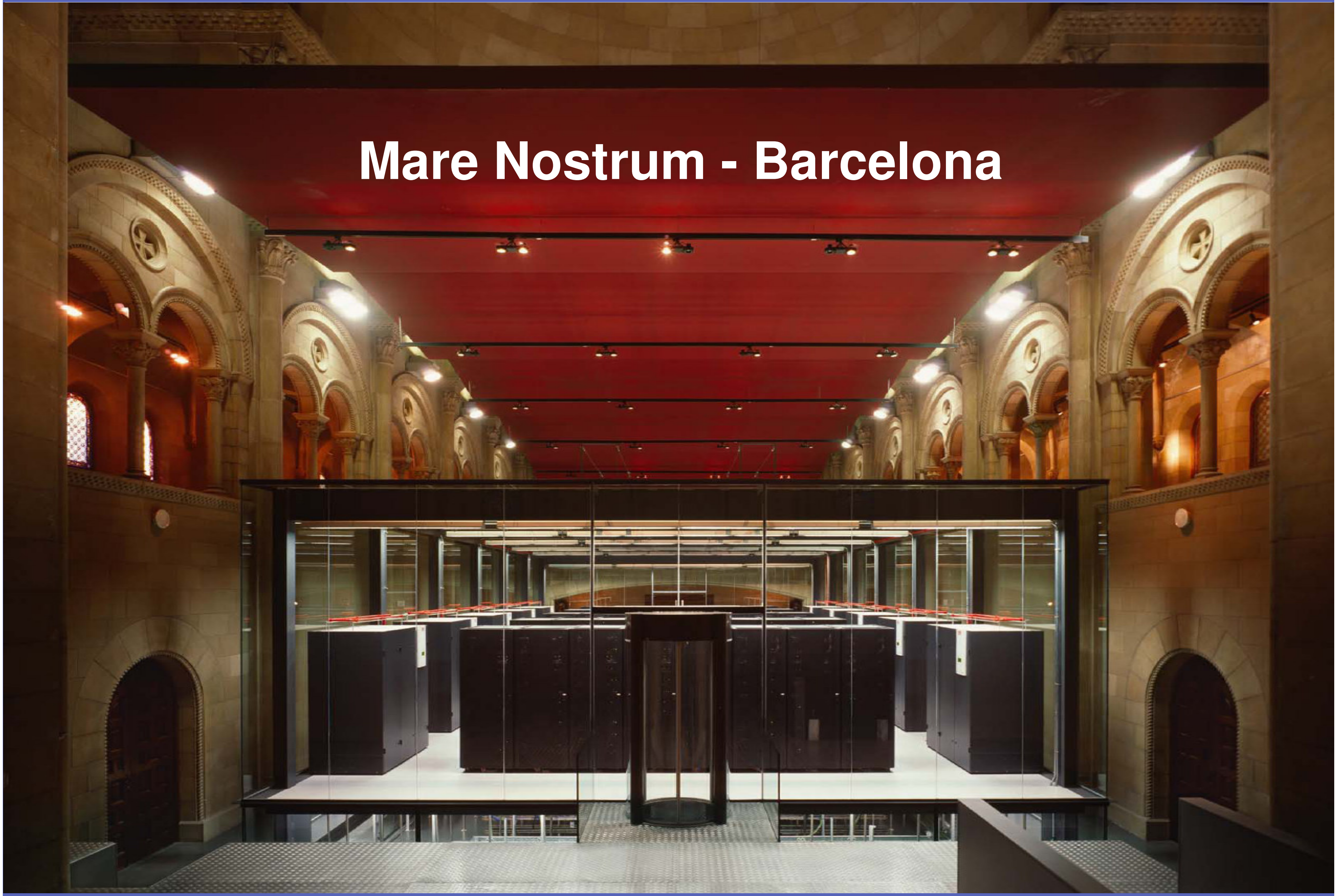
Grid / Workload Scheduling – IBM Tivoli Load Leveler, WebSphere XD

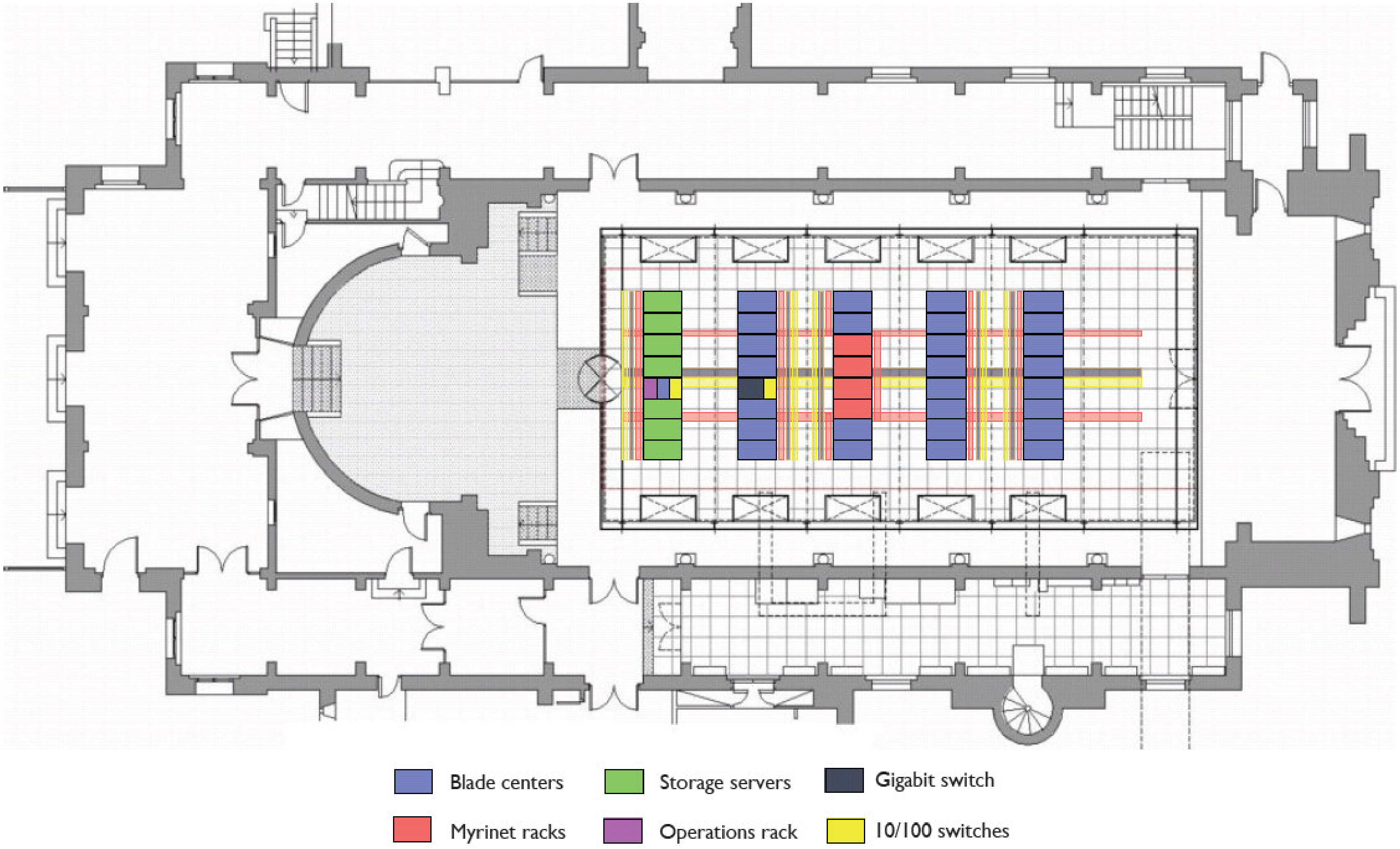
Data Cache – IBM Object Grid, IBM SolidDB, Informix IDS, DB2 PureXML

High Performance Parallel File System
IBM General Parallel File System

x86 / Cell / POWER
DCCoD
XIV Storage

Mare Nostrum - Barcelona




🟦 Blade centers
🟩 Storage servers
⬛ Gigabit switch
🟥 Myrinet racks
🟪 Operations rack
🟨 10/100 switches

The next wave – Application Optimised Systems

Processing

- IBM Blue Gene/P
- Cell Broadband Engine
- FPGAs / CPLDs /ASICs etc.
- Utility computing
- Computational appliances; e.g Azul Systems
- AGEIA's PhysX processor
- Google Enterprise Search appliances
- Graphics Processing Units (GPUs) – e.g. Nvidia

Storage appliances

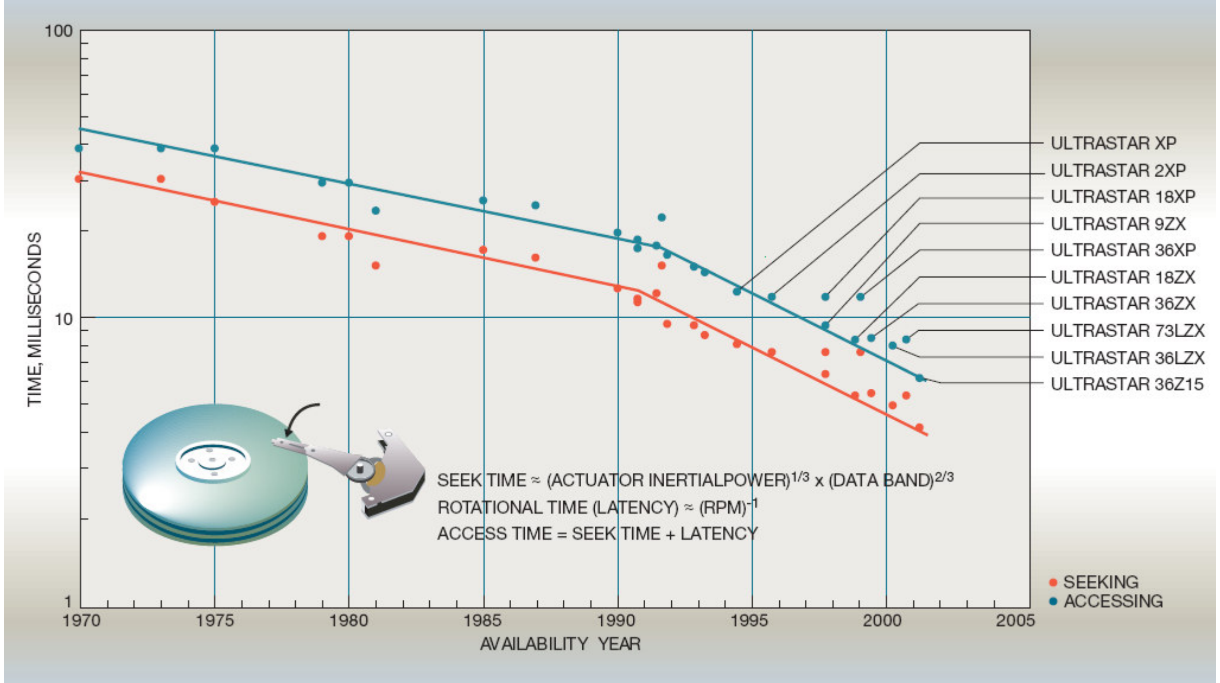
- Application-optimised Network-attached storage

Communication

- Network accelerators
- Protocol offload engines; e.g. DataPower XML accelerator
- Specialised interconnects

System performance has grown faster than disk access performance

Access Latency



Hard Drive Circa 1970

- HDD have provided capacity at a much greater rate than performance increases over the last 50 years.
  - HDD growth: 60+% since 1990
  - HDD access latency: <10% / y
- Systems have been optimized for an increasing disparity between computation and rotating disk performance
  - Chip-level performance growth 45% / y or more

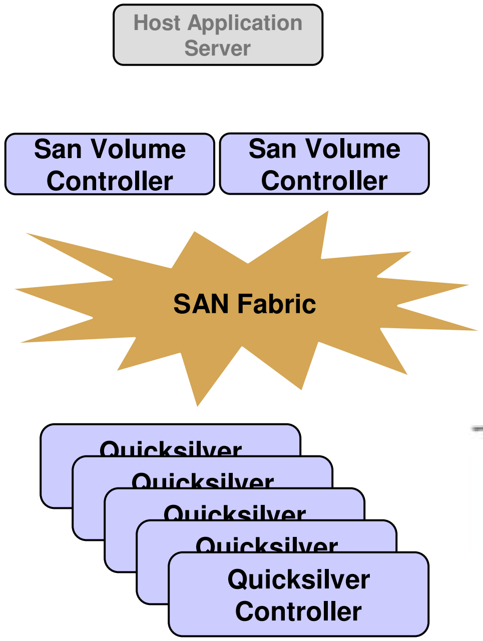
Thumb Drive

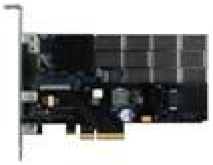
Quicksilver Flash Optimized Controller Prototype

- **Quicksilver is a Fibre-channel attached storage controller containing solid state storage devices.**
  - ►**SVC cluster provides vdisk provisioning and hot-swap management for the pool of solid state storage**

- **A cluster of SVC nodes and Quicksilver controllers achieved over 1 Million IOPs**
  - ►**(70/30 Read /Write mix 4K random I/O)**

- **[1] For comparison – the same 70/30 workload was performed on an 8 node SVC cluster with 1536 15k RPM HDD**
  - ►**This SVC cluster configuration used in the published SPC-1 benchmark.**

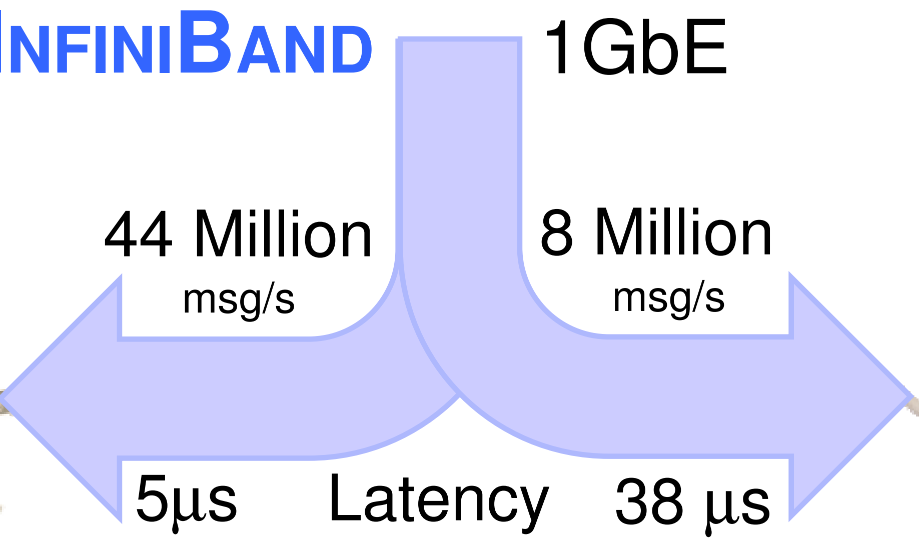




Integration of SSD's in p570 16 way Configuration

- Same throughput performance with equal or better response time
- Reduces DRAM by 50% (from 512GB – 256GB
- Reduces hard disk drives by 50% (from 1,667 to 850)
- System cost, floor space and energy savings (30-40%)
- Slight increase in CPU Utilization from 77% to 86%

Network and Middleware Transport



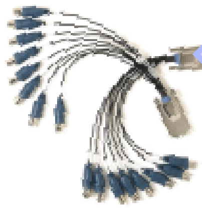
INFINIBAND

1GbE

44 Million
msg/s

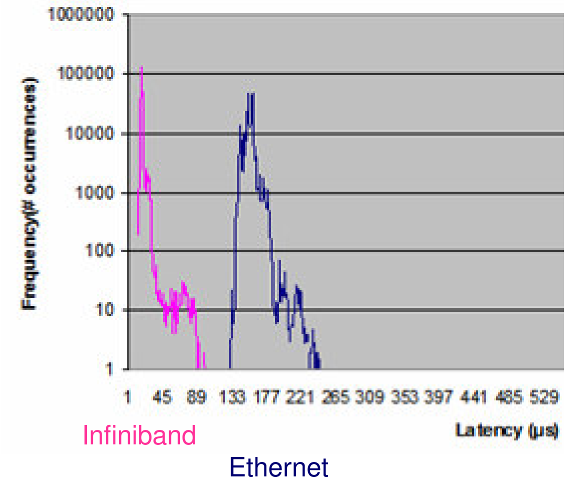
8 Million
msg/s

5µs Latency 38 µs



Message size in bytes	QDR 6 stream	
	msgs /sec	Mbits /sec
12	43.9M	4.3K
45	13.3M	4.8K
120	5M	4.9K
1.2K	428K	4.0K
12K	49.8K	4.8K
120K	5K	4.9K

Latency
for 1Kb 50Km/s



Case Study

Algorithmic Trading

MQ Low Latency Messaging

Software Linux kernel 2.6.29

Hardware Voltaire InfiniBand

Core i7 Blade Servers

Nagios

What will you create?