BIG DATA

**The Petabyte Age:
More Isn't Just More —
More Is Different**

wired magazine Issue 16/.07

Martien Ouwens

Datacenter Solutions Architect

Sun Microsystems

Big Data, The Petabyte Age:

The biggest challenge of the Petabyte Age won't be storing all that data, it'll be figuring out how to make sense of it

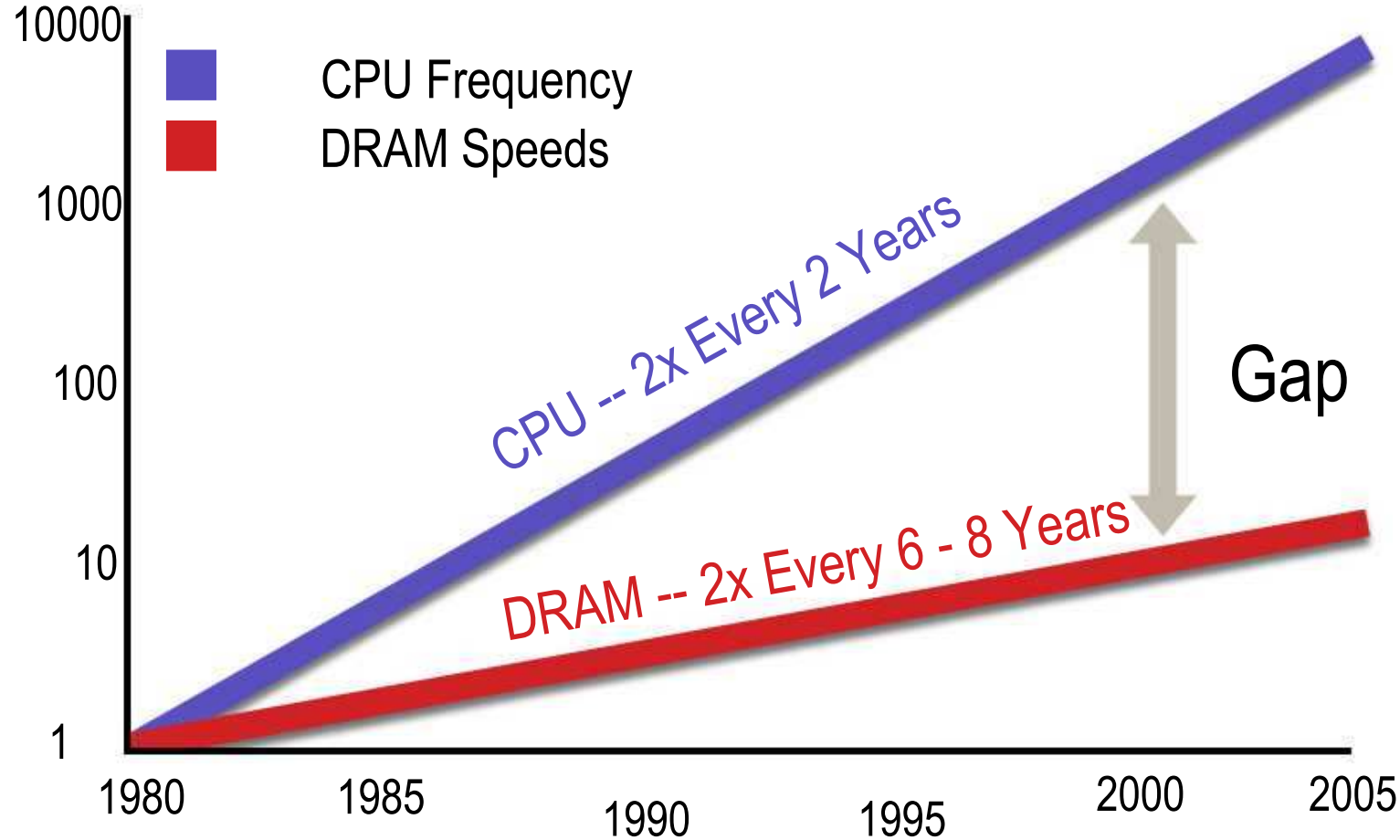**wired magazine Issue 16/.07**

Big Data Key Trends

1 TERABYTE

A $200 HARD DRIVE
THAT HOLDS
260,000 SONGS.

20 TERABYTE

PHOTOS UPLOADED TO
FACEBOOK EACH MONTH

120 TERABYTE

ALL THE DATA
AND IMAGES
COLLECTED BY
THE HUBBLE
SPACE TELESCOPE.

330 TERABYTE

DATA THAT
THE LARGE HADRON
COLLIDER WILL
PRODUCE EACH WEEK.

460 TERABYTE

ALL THE DIGITAL
WEATHER
DATA COMPILED
BY THE NATIONAL
CLIMATIC DATA
CENTER.

530 TERABYTE

ALL THE VIDEOS
ON YOUTUBE.

600 TERABYTE

ANCESTRY.COM'S
GENEALOGY
DATABASE (INCLUDES
ALL U.S. CENSUS
RECORDS 1790-2000).

1 PETABYTE

DATA PROCESSED
BY GOOGLE'S
SERVERS EVERY
72 MINUTES.

"Hitting walls" in Processor Design

- Clock frequency
  - frequency increases tapering off, in new semiconductor processes
  - high frequencies => *power* issues

- Memory latency (not instruction execution speed) dominating most application times

- Processor designs for high single-thread performance are becoming *highly* complex, therefore:
  - expense and/or time-to-market suffer
  - verification increasingly difficult
  - more complexity => more circuitry => increased power ... for diminishing performance returns

Memory Bottleneck

Relative

Performance

Legend:
- CPU Frequency
- DRAM Speeds

CPU -- 2x Every 2 Years

DRAM -- 2x Every 6 - 8 Years

Gap

Y-axis: 10000, 1000, 100, 10, 1

X-axis: 1980, 1985, 1990, 1995, 2000, 2005

Source: Sun World Wide Analyst Conference Feb. 25, 2003
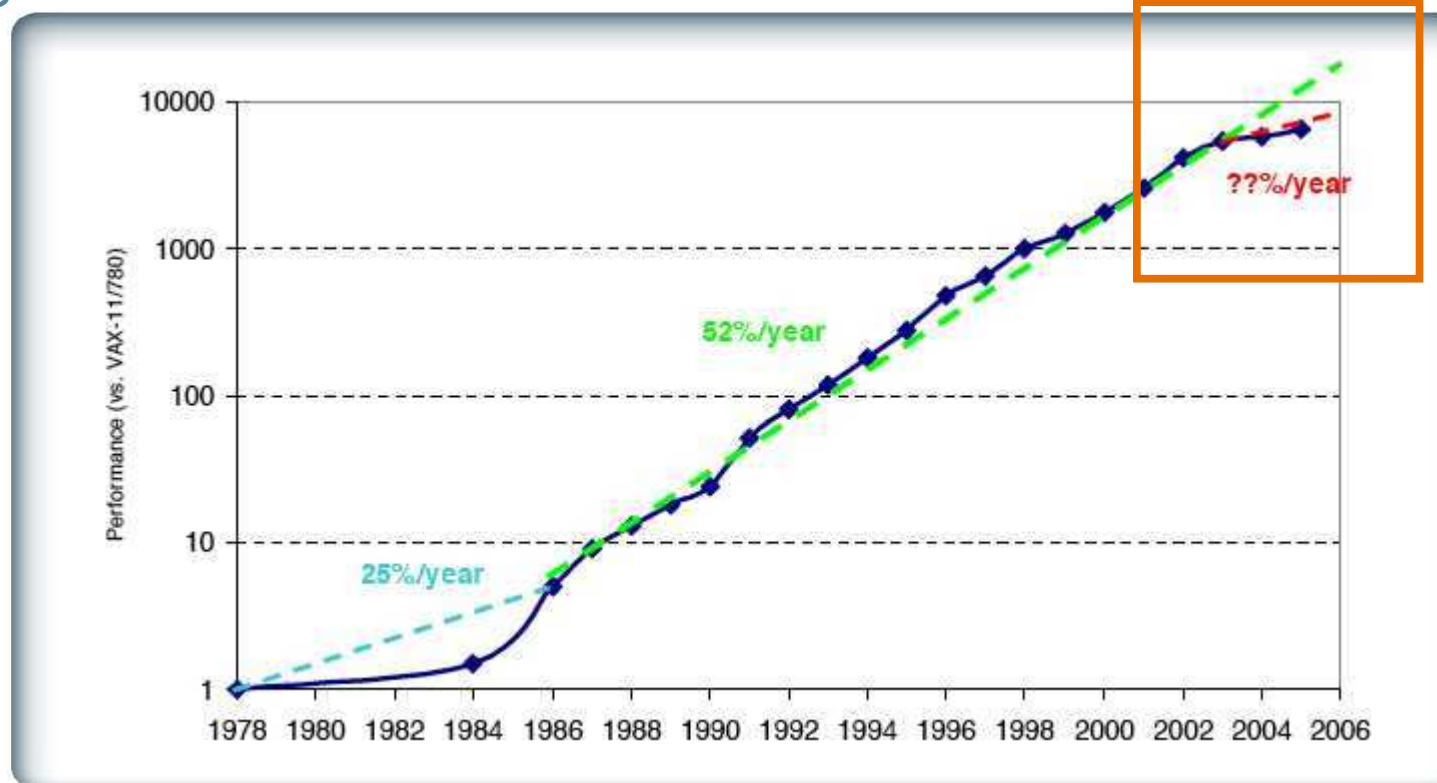
How We Mask Memory Latency Today
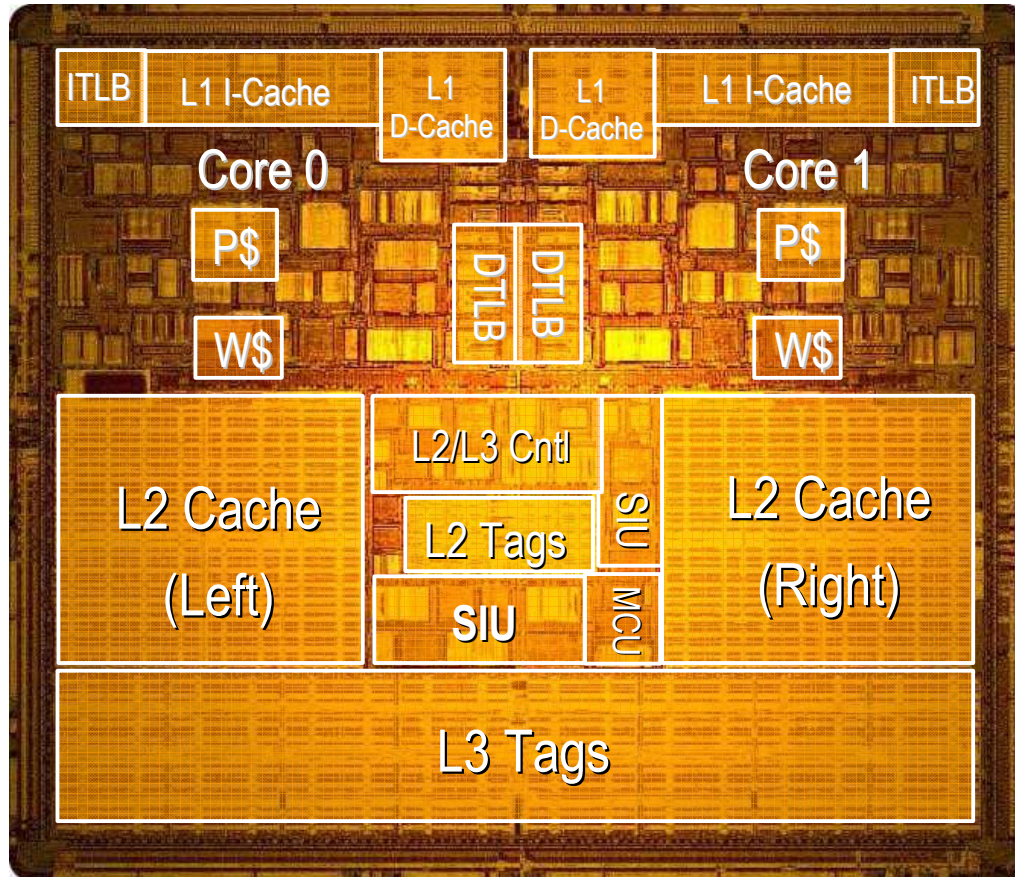
Great Big Caches

- of many different kinds

- that only mask memory latency

- and execute no code

- accessed one cache line at a time

- but require power and cooling to all lines all the time

Cache Logic Accounts for About 75% of the Chip Area

"Power Wall + Memory Wall + ILP Wall = Brick Wall"

"In 2006, performance is a factor of three below the traditional doubling every 18 months that we enjoyed between 1986 and 2002.
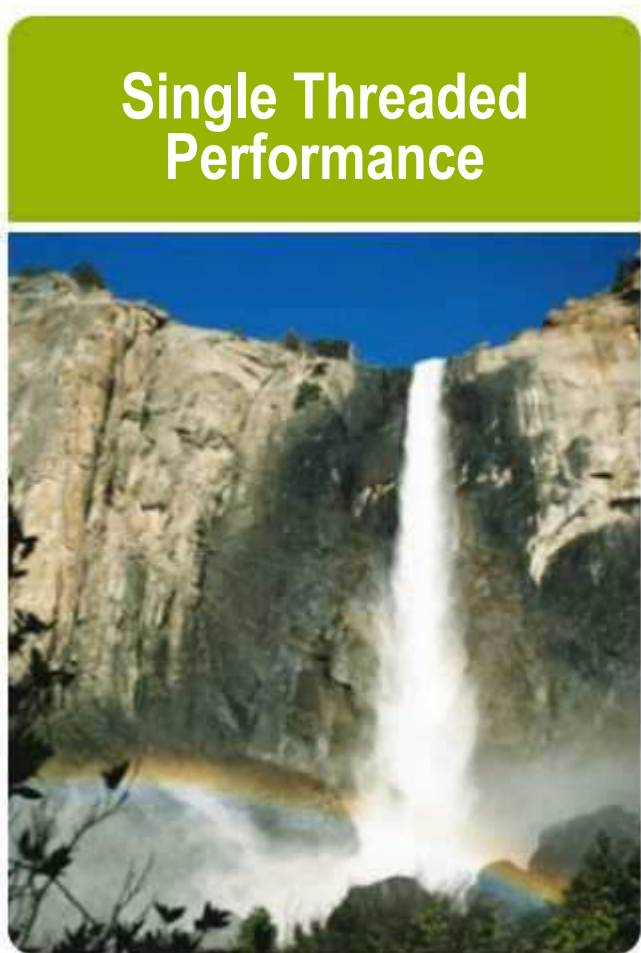
The doubling of uniprocessor performance may now take *5 years*." *

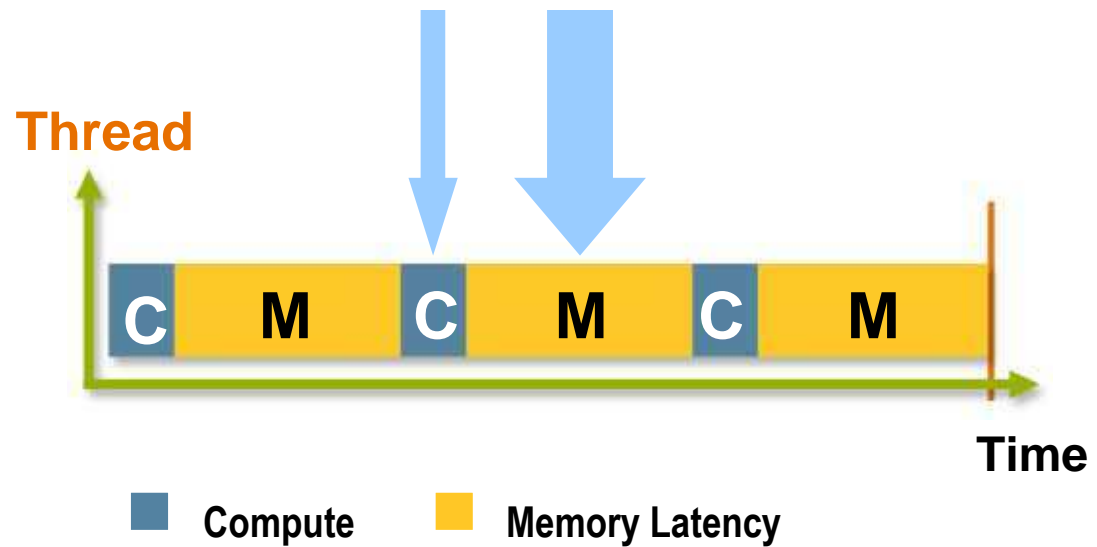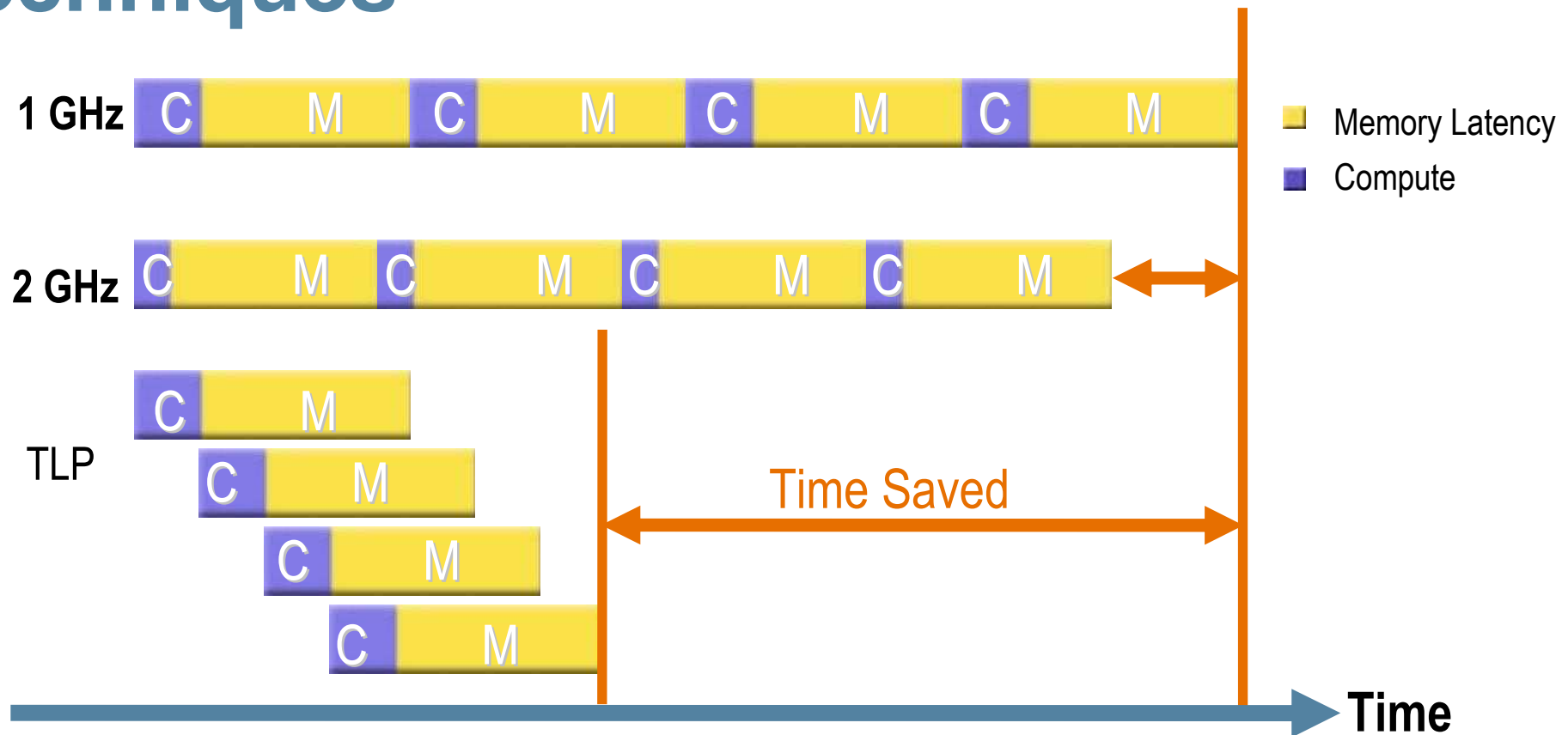* "The Landscape of Parallel Computing Research: A View from Berkeley", December 2006 (emphasis added)
http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.html

Single Threading
Up to 85% Cycles Spent Waiting for Memory

Hurry Up and WAIT

Single Threaded Performance

SPARC US-IV+	**25%**	**75%**
Intel	**15%**	**85%**

Thread

C	M	C	M	C	M

Time

■ **Compute**   ■ **Memory Latency**

Comparing Modern CPU Design Techniques

ILP Offers Limited Headroom
TLP Provides Greater Performance Efficiency

Chip Multi-Threading (CMT)
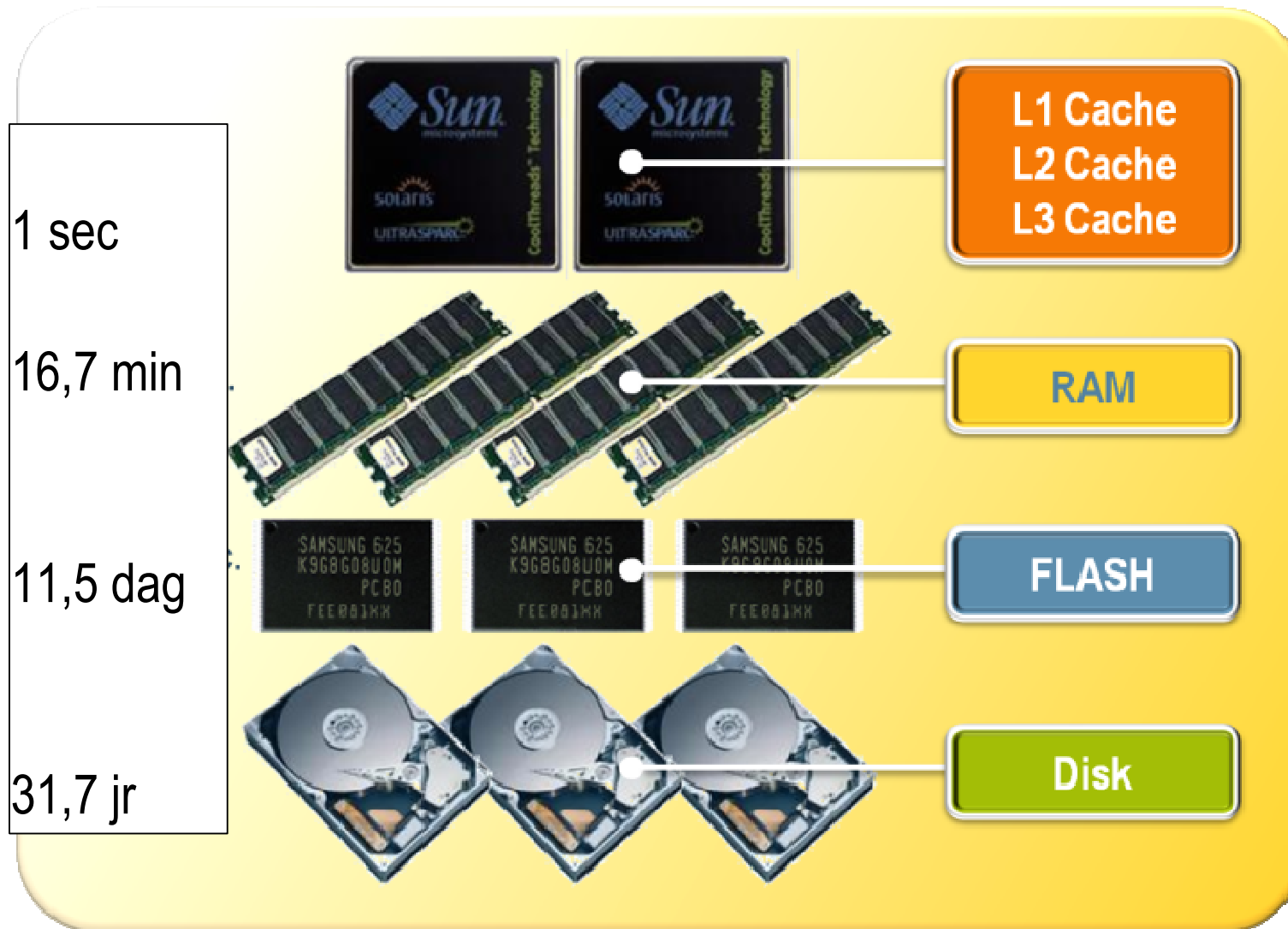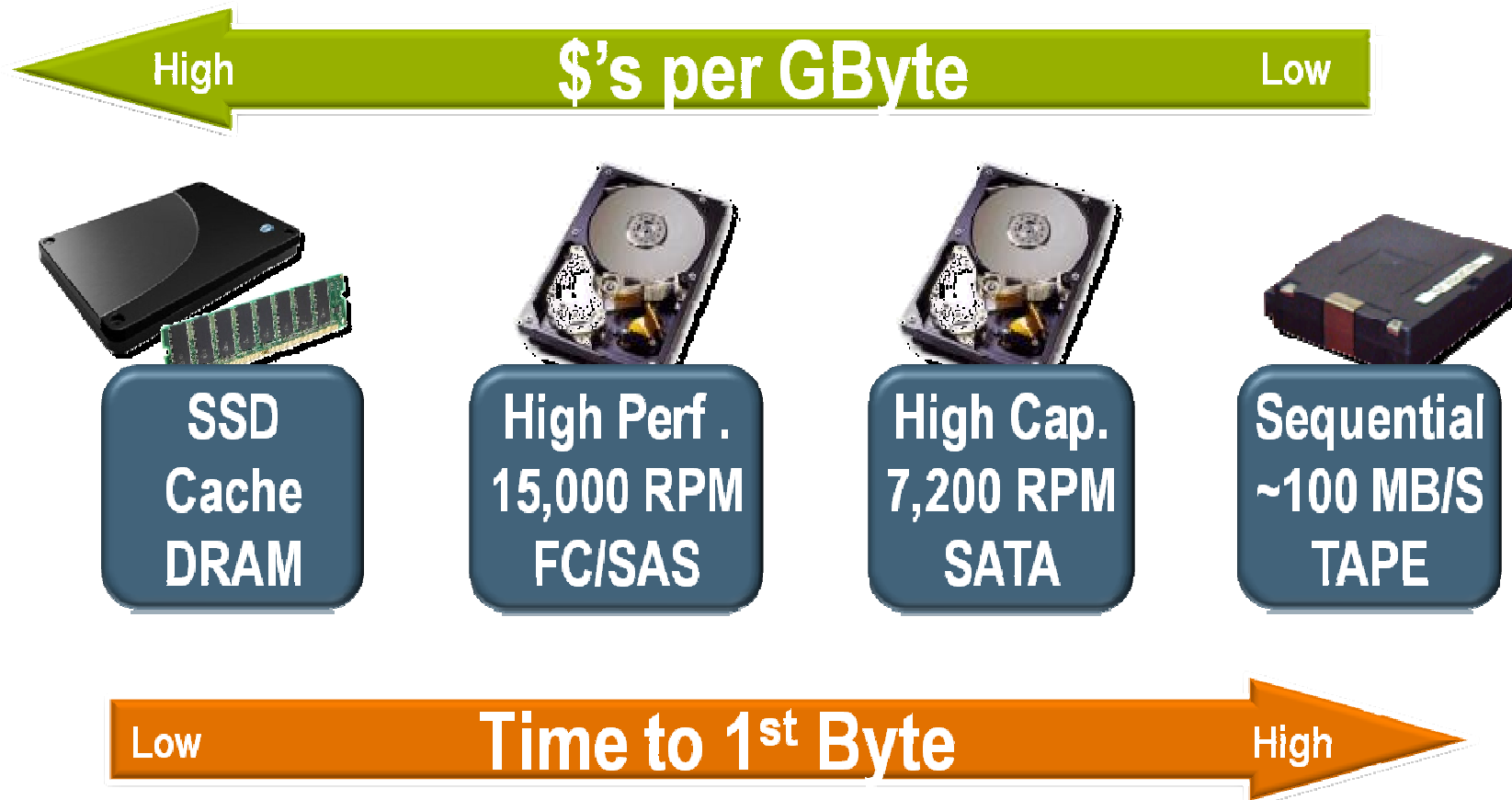## Utilization: Up to 85%*

UltraSPARC-T1   85% 15%

Compute   Latency

Core-1 Thread 4 / Thread 3 / Thread 2 / Thread 1

Core-2 Thread 4 / Thread 3 / Thread 2 / Thread 1

Core-3 Thread 4 / Thread 3 / Thread 2 / Thread 1

Core-4 Thread 4 / Thread 3 / Thread 2 / Thread 1

Core-5 Thread 4 / Thread 3 / Thread 2 / Thread 1

Core-6 Thread 4 / Thread 3 / Thread 2 / Thread 1

Core-7 Thread 4 / Thread 3 / Thread 2 / Thread 1

Core-8

Chip Multi-threaded (CMT) Performance

Time

* based on example of UltraSPARC T1

10

Enterprise Flash – Relative Perf

1 sec	L1 Cache / L2 Cache / L3 Cache
16,7 min	RAM
11,5 dag	FLASH
31,7 jr	Disk

Data Retention Comparison

	DRAM	SSD	HDD
Budgetary Cost	$100/GB	$35/GB	$5/GB
Power Consumption	463 Watts	2.5 Watts	12 Watts
Random I/O	1,000,000 IOPS	50,000 IOPS	350 IOPS

ZFS Turbo Charges Applications

## Hybrid Storage Pool Data Management on Solaris Systems, Open Storage
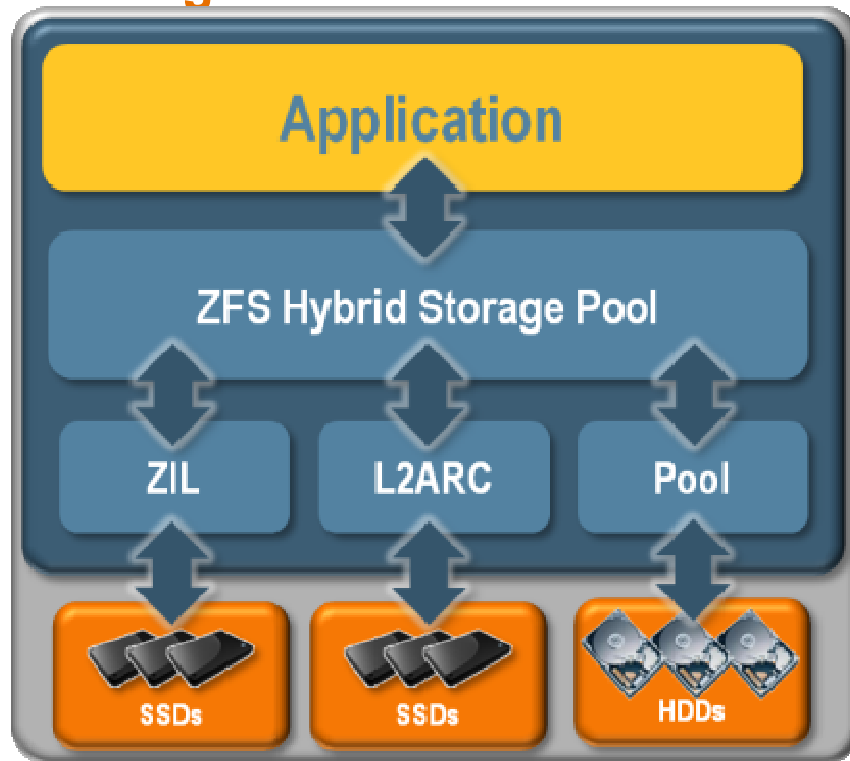
ZFS automatically:

- Writes new data to very fast SSD pool (ZIL)

- Determines data access patterns and stores frequently accessed data in the L2ARC

- Bundles IO into sequential lazy writes for more efficient use of low cost mechanical disks

Unique Capabilities from Sun
## Sun Storage 7000 Unified Storage Systems
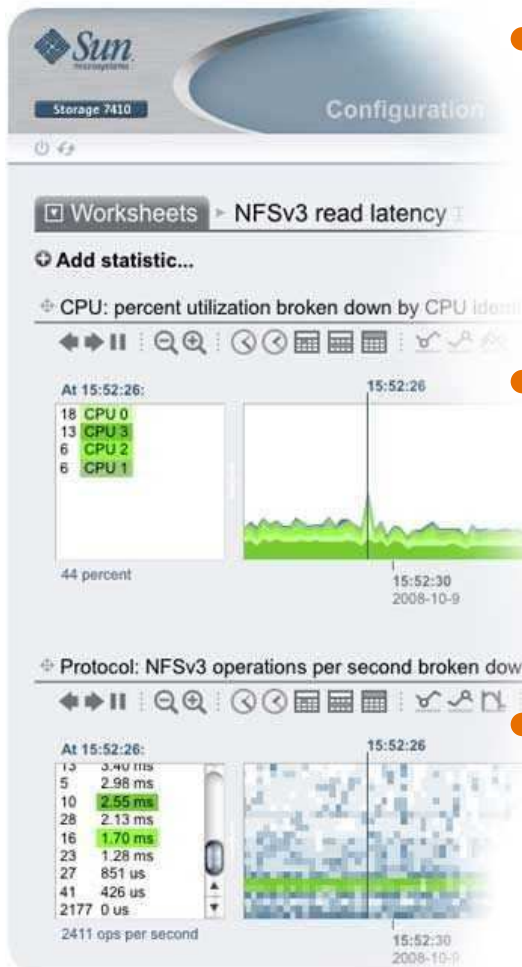
- **Radical simplification**
  - > Ready-to-go appliance – installs in minutes
  - > Extensive analytics and drill-down for rapid management of performance challenges

- **Improved latency and performance**
  - > Integrated flash drives
  - > Hybrid Storage Pools automates data placement on the best media

- **Compelling value through standards**
  - > Industry standard hardware and open source software vs. proprietary solution

Sun Oracle Database Machine

- Hardware by Sun, software by Oracle – Exadata Version 2
- Enhancements to Exadata Version 1 offering extreme performance
  - > Faster servers, storage, and network interconnects
  - > Flash acceleration
  - > Speedup at Oracle level of 20 times processing IOPS
  - > Extends Exadata to provide acceleration for OLTP
- Database aware storage
  - > Smart scans
  - > Storage indexes – eliminates a good deal of I/Os
- Massively parallel architecture
- Dynamic linear scaling
- Offering mission critical availability and protection of data

What Does Sun Oracle Database Machine Address?
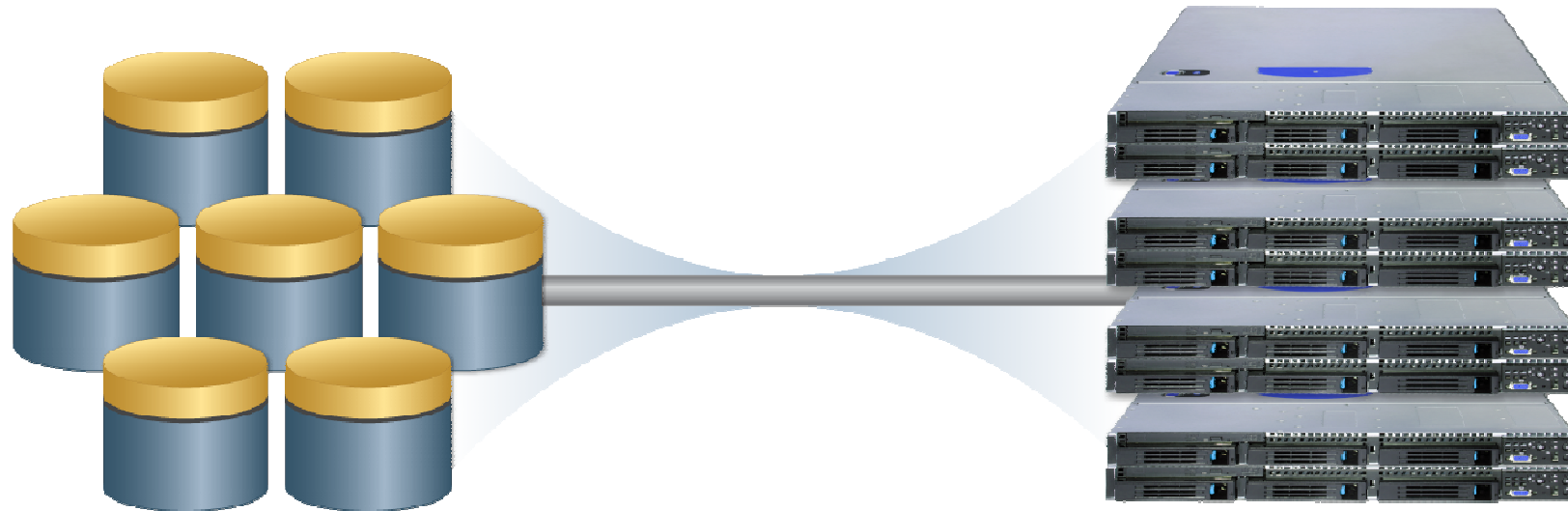
- Current database deployments often have bottlenecks limiting the movement of data from disks to servers
  - > Storage Array internal bottlenecks on processors and Fibre Channel Architecture
  - > Limited Fibre Channel host bus adapters in servers
  - > Under configured and complex SANs
- Pipes between disks and servers are 10x to 100x too slow for data size

Sun Oracle Database Machine Solves the Bottleneck by
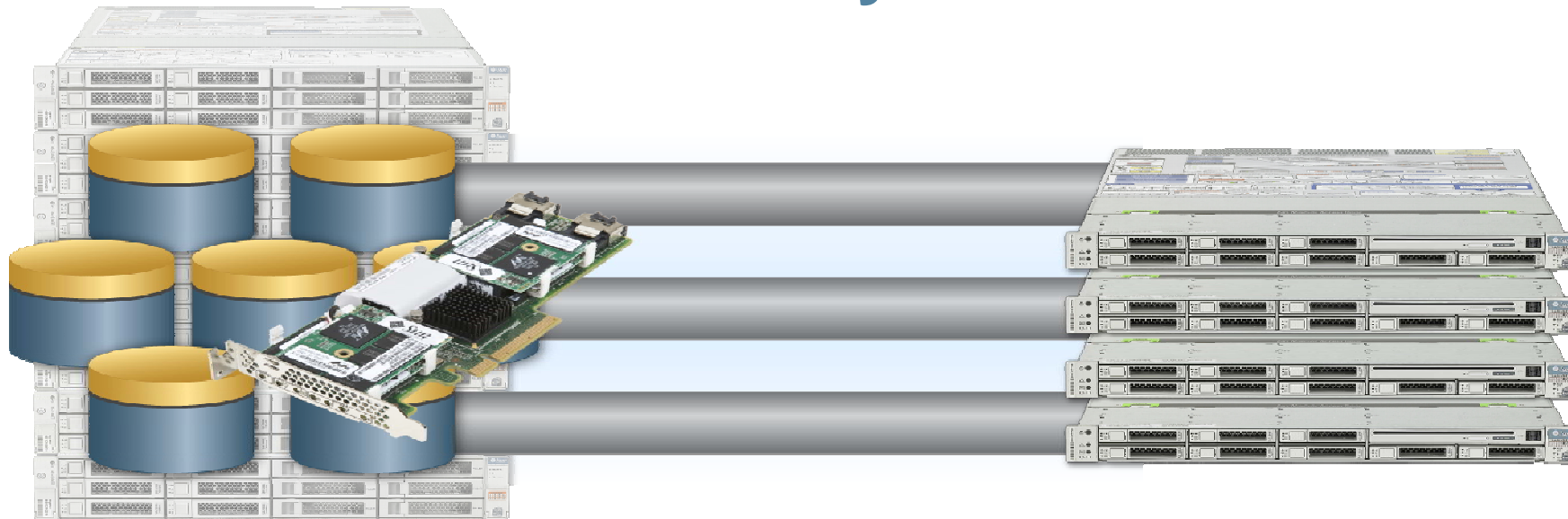

- Adding more pipes – Massively parallel architecture

- Make the pipes wider – 5X faster than conventional storage

- Ship less data through the pipes – Process data in storage

- Simplify the storage offering – eliminate the need of a SAN

- Allow for Ease of Growth

- Provide a Flash Cache for Rapid Response

HPC Market Trends

- It's a growing market, outpacing growth of many others
  - > $10B in 2008
  - > CAGR of 3.1% for 2007 to 2012*

- Clustering is now the dominant architecture

- HPC technologies now within reach of all organizations

HPC is Underserved by Moore's Law
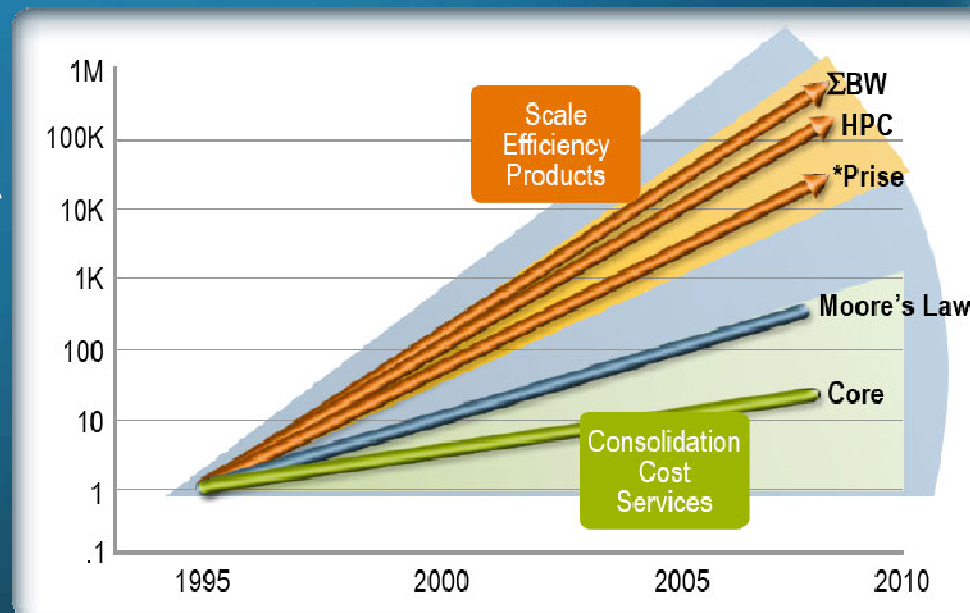
* Source IDC (January 2009)

Sun Constellation System Open Petascale Architecture

Eco-Efficient Building Blocks

Compute	Networking	Storage	Software
			DEVELOPER TOOLS
			GRID ENGINE
			PROVISIONING

Ultra-Dense Blade Platform

- Fastest processors: AMD Opteron, Intel Xeon
- Highest compute density
- Fastest host channel adaptor

Ultra-Dense Switch Solution

- 648 port InfiniBand switch
- Unrivaled cable simplification
- Most economical InfiniBand cost/port

Ultra-Dense Storage Solution

- Most economical and scalable parallel file system building block
- Up to 48 TB in 4RU
- Up to 2TB of SSD
- Direct cabling to IB switch

Comprehensive Software Stack

- Integrated developer tools
- Integrated Grid Engine infrastructure
- Provisioning, monitoring, patching
- Simplified inventory management

Sun Constellation System Open Petascale Architecture

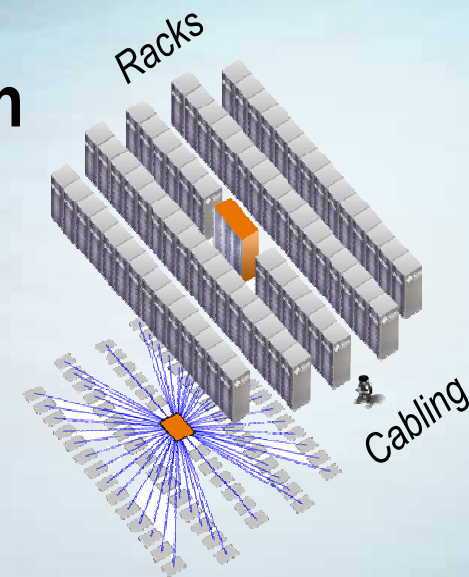## Radical Simplicity, Faster Time to Deployment

Competitors Compute Clusters

Racks

Leaf Switches

Core Switches
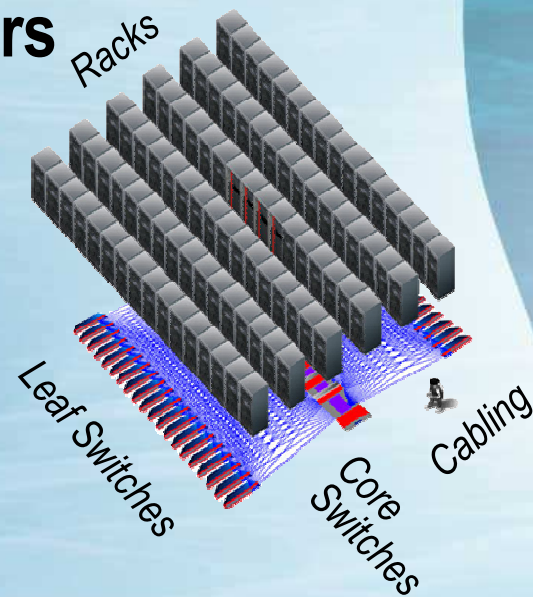
Cabling

Competitors Cabling Infrastructure

- **Alternative Open Standards Fabric**
- 300 switching elements
- 6912 cables
- 92 racks

Sun Constellation System Cluster

Racks

Cabling

Reduced Cabling

- 1 switching element **300:1 reduction**
- 1152 cables **6:1 reduction**
- 74 racks **20% smaller footprint**

Compute Power —
579 TFLOPS Aggregate Peak

- 3,936 Sun four-socket, quad-core nodes
- 15,744 AMD Opteron processors
- Quad-core, four flops/cycle (dual pipelines)

Memory

- 2 GB/core, 32 GB/node, 125 TB total
- 132 GB/s aggregate bandwidth

Disk Subsystem

- 72 Sun x4500 "Thumper" I/O servers, 24TB each
- 1.7 Petabyte total raw storage
- Aggregate bandwidth ~32 GB/sec

InfiniBand Interconnect

Massive Low Latency Engine

The HPC Data I/O Challenge

- Compute power is scaling rapidly
  - > Larger compute clusters
  - > More sockets per server
  - > More cores per CPU

- Compute advances far outpace those in I/O
  - > Storage latency 12 orders of magnitude higher than compute
  - > Compute Side developed Parallel Approaches sooner

- Need a better approach to feed compute
  - > Cluster performance is often limited by data I/O
  - > Need a parallel I/O approach to help balance things

Parallel Storage
Sun Storage Cluster
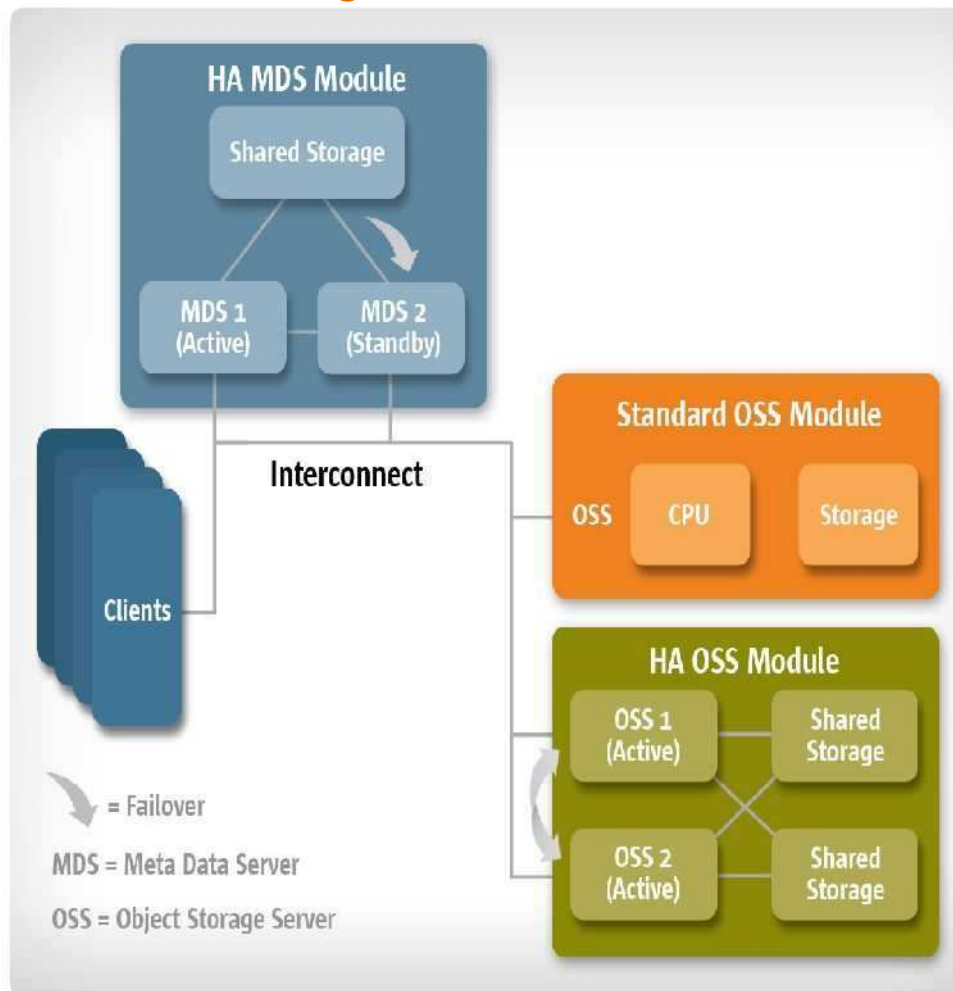
- Lustre and Open Storage providing a wide range of Scaling and Performance for a wide range of cluster sizes

- Breakthrough economics
  - > Up to 50% cost savings vs. competitors

- Simplified deployment and management of Lustre storage

Unique Capabilities from Sun lustre

Sun Storage Cluster

- High performance and broad scaling
  - > From a few to over 100 GB/second
  - > From dozens to thousands of nodes

- Breakthrough economics, up to 50% savings
  - > Built with standardized hardware and open source software vs. proprietary products

- Simplifying Lustre for the mainstream
  - > Pre-defined modules speed deployment of parallel file system, available factory integrated
  - > Ease of use improvements – patchless clients, integrated driver stack, LNET self-test, mount-based cluster configuration, many more

Lustre in Action

TACC Ranger
1.73 PB storage, 35GB/s I/O throughput, 3,936 quad-core clients

Sandia Red Storm
340 TB Storage, 50GB/s I/O throughput, 25,000 clients

Framestore
"The Tale of Despereaux"
200TB Lustre file system – averaged 1.2GB/s in sustained reads, peaks of 3GB/s, 5TB data generated per night from cluster of 4,000 cores interfacing with Lustre file system

German Climate Research Data Centre (DKRZ)
272 TB, Lustre and Storage Archive Manager, 256 nodes with 1024 cores
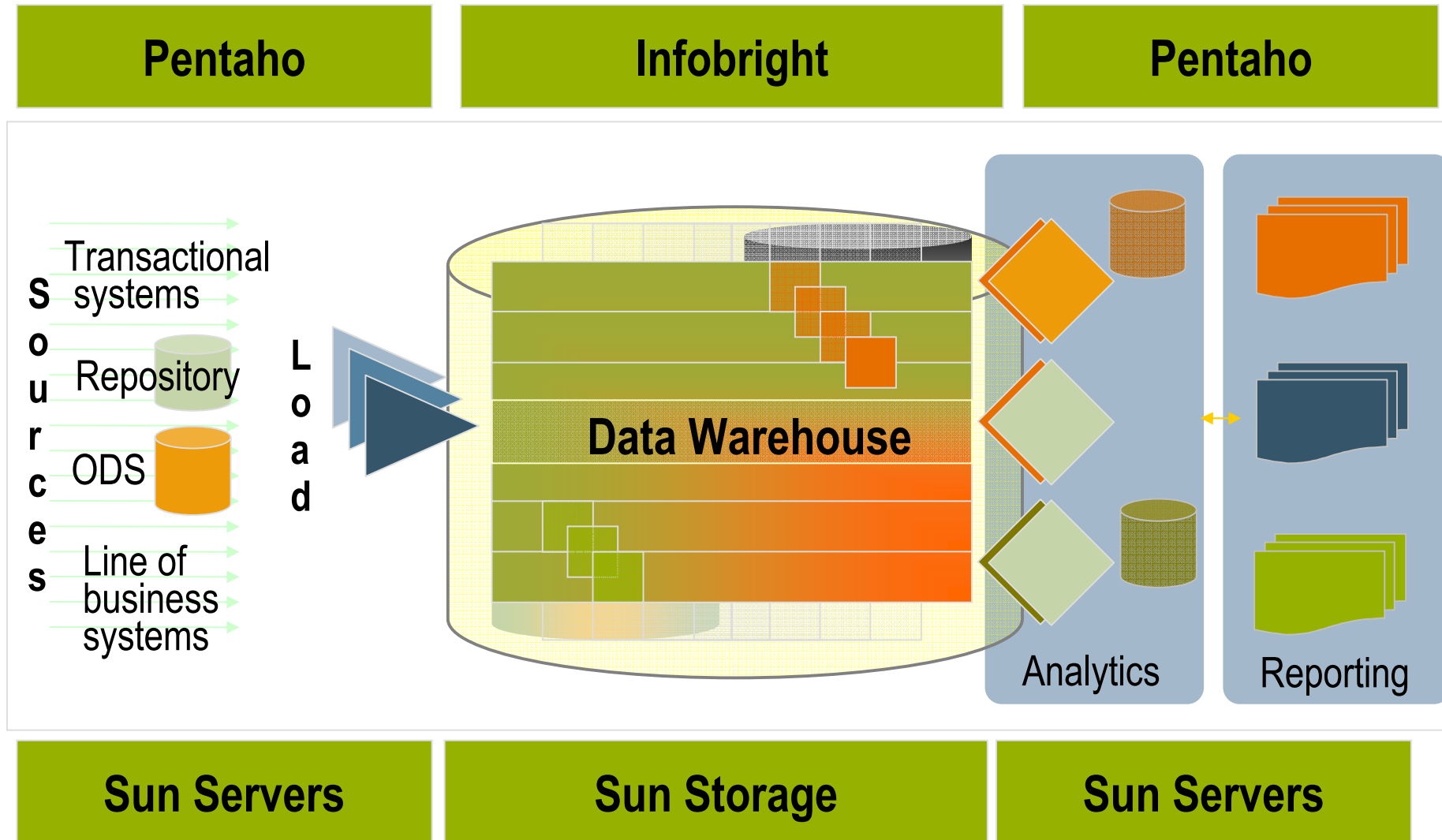
Big Data Key Trends
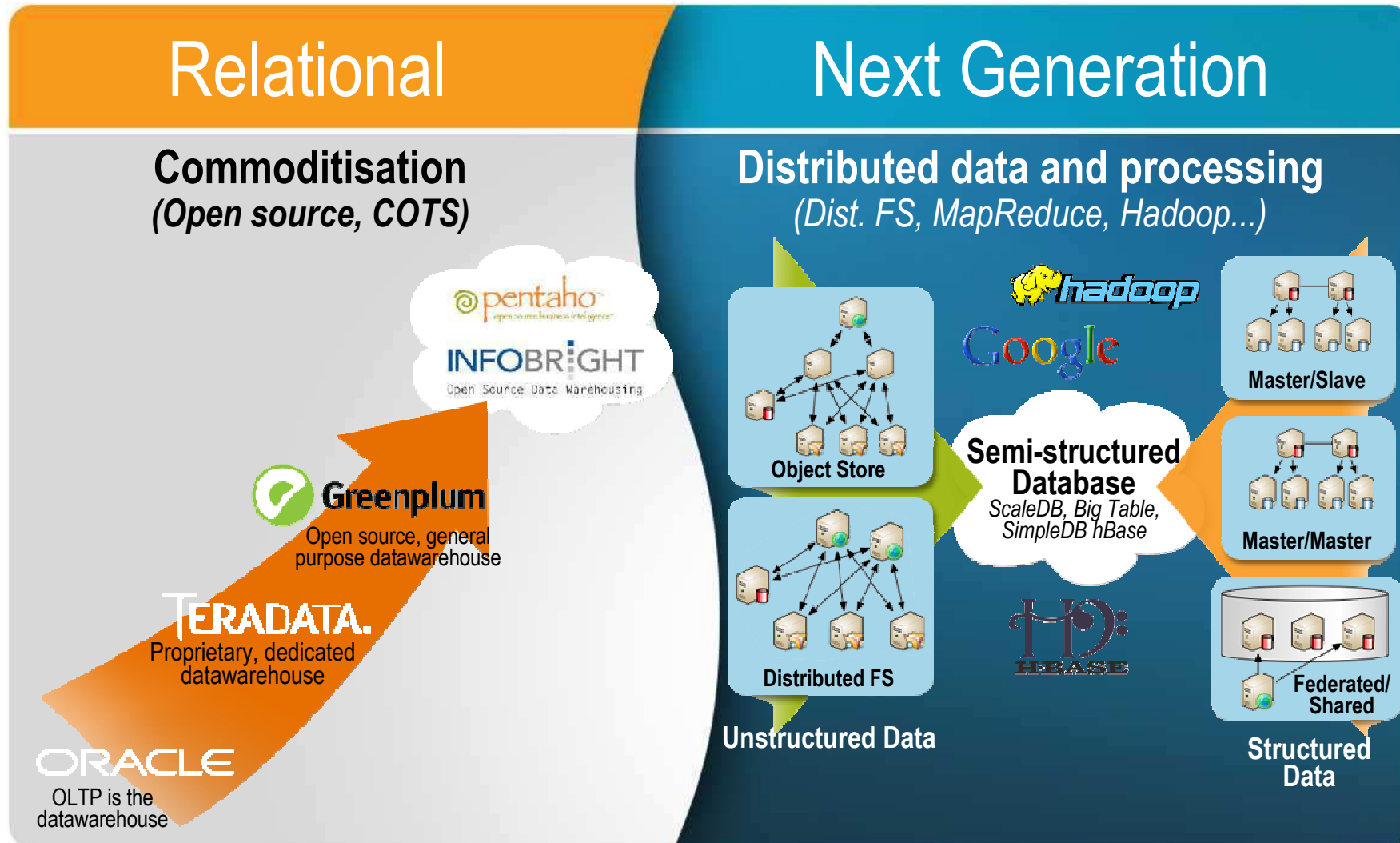
Big Data

Martien Ouwens
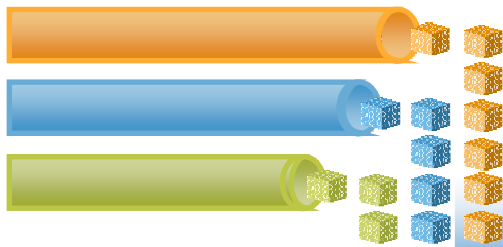Datacenter Solutions Architect
martien.ouwens@sun.com

Sun Microsystems Nederland B.V.

Typical BI/DW Components

Column-Orientation
## Exemplified by InfoBright's approach

Column Orientation

Data Packs – data stored in manageably sized, highly **compressed data** packs

Knowledge Grid – statistics and metadata "describing" the super-compressed data which is automatically updated as data packs are created or updated

Infobright Optimizer iterates over the Knowledge Grid; only the data packs needed to resolve the query are decompressed
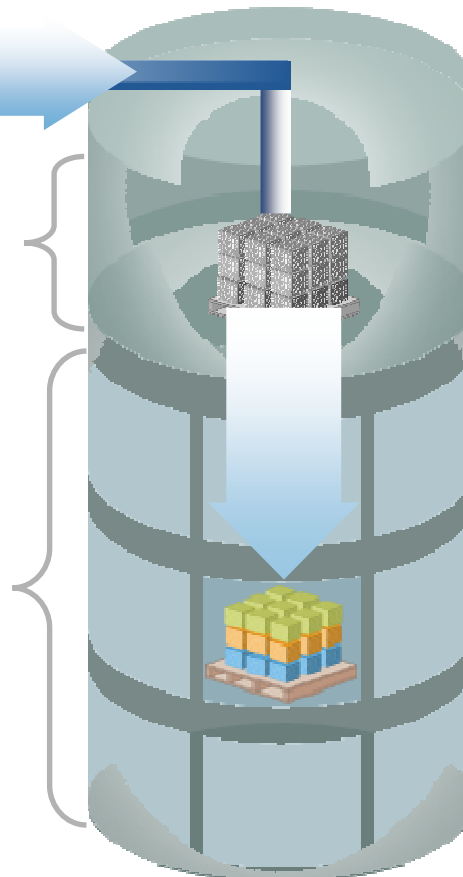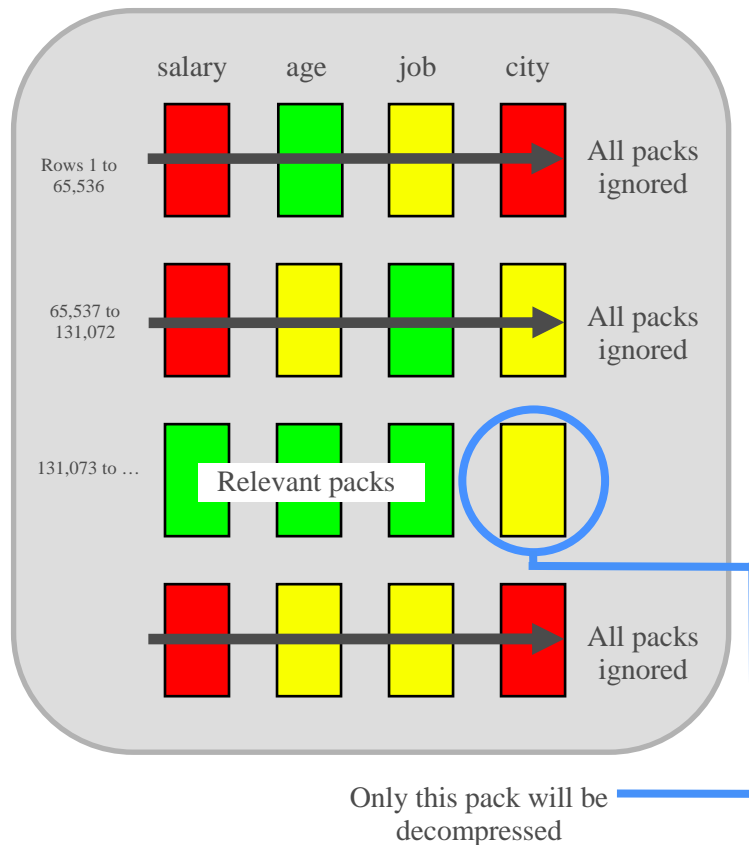
Smarter architecture

- Load data and go
- No indexes or partitions to build and maintain, no tuning
- Super-compact data foot- print leverages off-the-shelf hardware
- MySQL wrapper connects to BI and ETL tools

30

InfoBright Query Execution

`employees` records stored in Data
Packs of 64k elements each

SELECT COUNT(*) FROM employees
 WHERE salary > 50000
 AND age < 65
 AND job = 'Shipping'
 AND city = 'TORONTO';

	salary	age	job	city	
Rows 1 to 65,536					All packs ignored
65,537 to 131,072					All packs ignored
131,073 to …	Relevant packs				
					All packs ignored

Only this pack will be decompressed

Using the Knowledge Grid, we verify, which packs are

• irrelevant (**no** values found that fit SELECT criteria)

• relevant (**all** values fit)

• suspect (**some** values fit)

Any row containing 1 or more irrelevant DP's is ignored.

Since we are looking for COUNT(*) in this case we do not need decompress relevant packs. COUNT information is kept in the Knowledge Grid in the Data Pack Nodes.

The suspect packs must be decompressed for detailed evaluation.