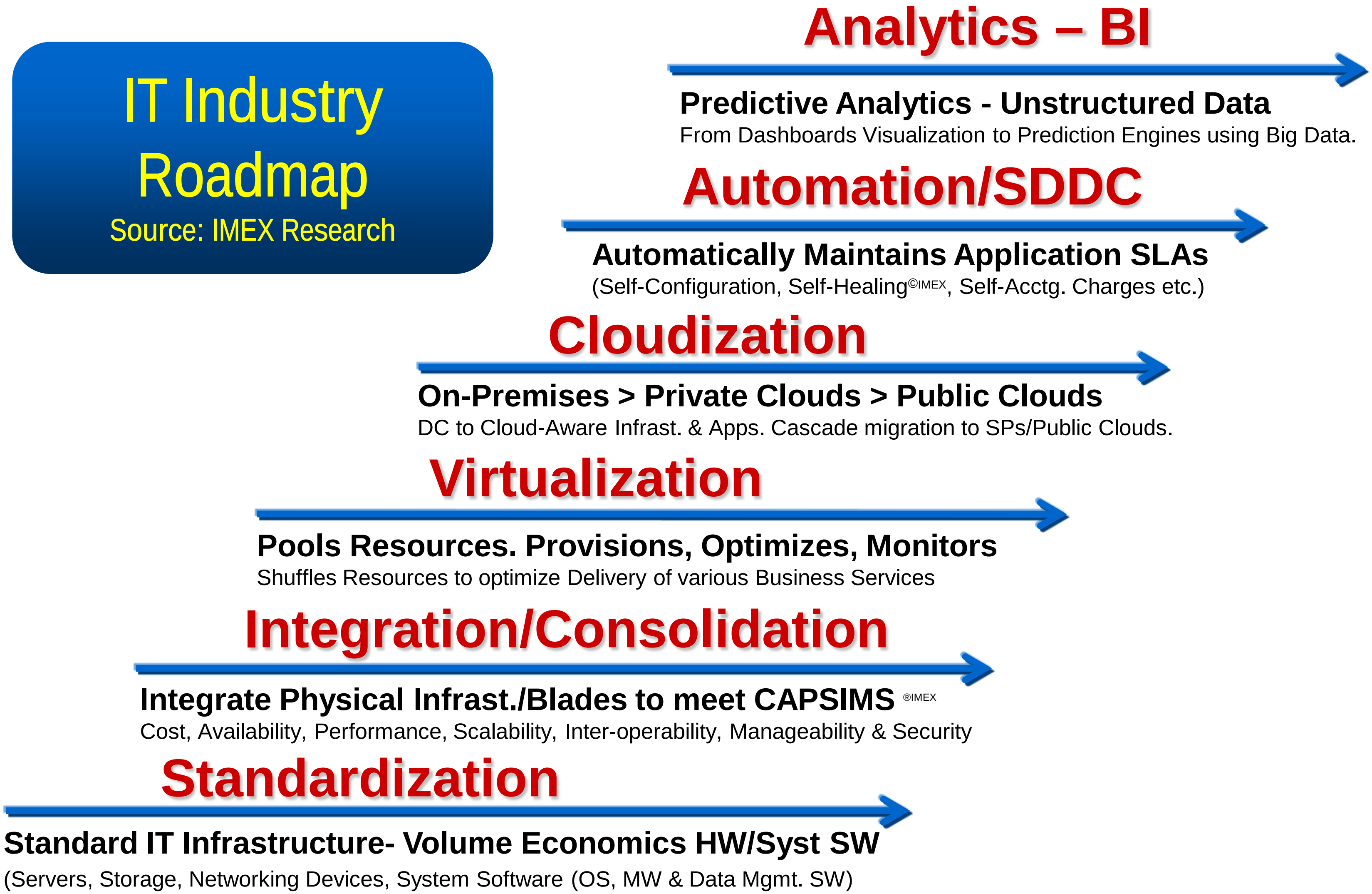
System Architecture
for
In-Memory Database

Anil Vasudeva
President & Chief Analyst
imex@imexresearch.com
408-268-0800

IMEX
RESEARCH.COM

IT Industry Dynamics - Roadmap



IT Industry Roadmap

Source: IMEX Research

Analytics – BI

Predictive Analytics - Unstructured Data

From Dashboards Visualization to Prediction Engines using Big Data.

Automation/SDDC

Automatically Maintains Application SLAs

(Self-Configuration, Self-Healing$^{\text{©IMEX}}$, Self-Acctg. Charges etc.)

Cloudization

On-Premises > Private Clouds > Public Clouds

DC to Cloud-Aware Infrast. & Apps. Cascade migration to SPs/Public Clouds.

Virtualization

Pools Resources. Provisions, Optimizes, Monitors

Shuffles Resources to optimize Delivery of various Business Services

Integration/Consolidation

Integrate Physical Infrast./Blades to meet CAPSIMS $^{\text{®IMEX}}$

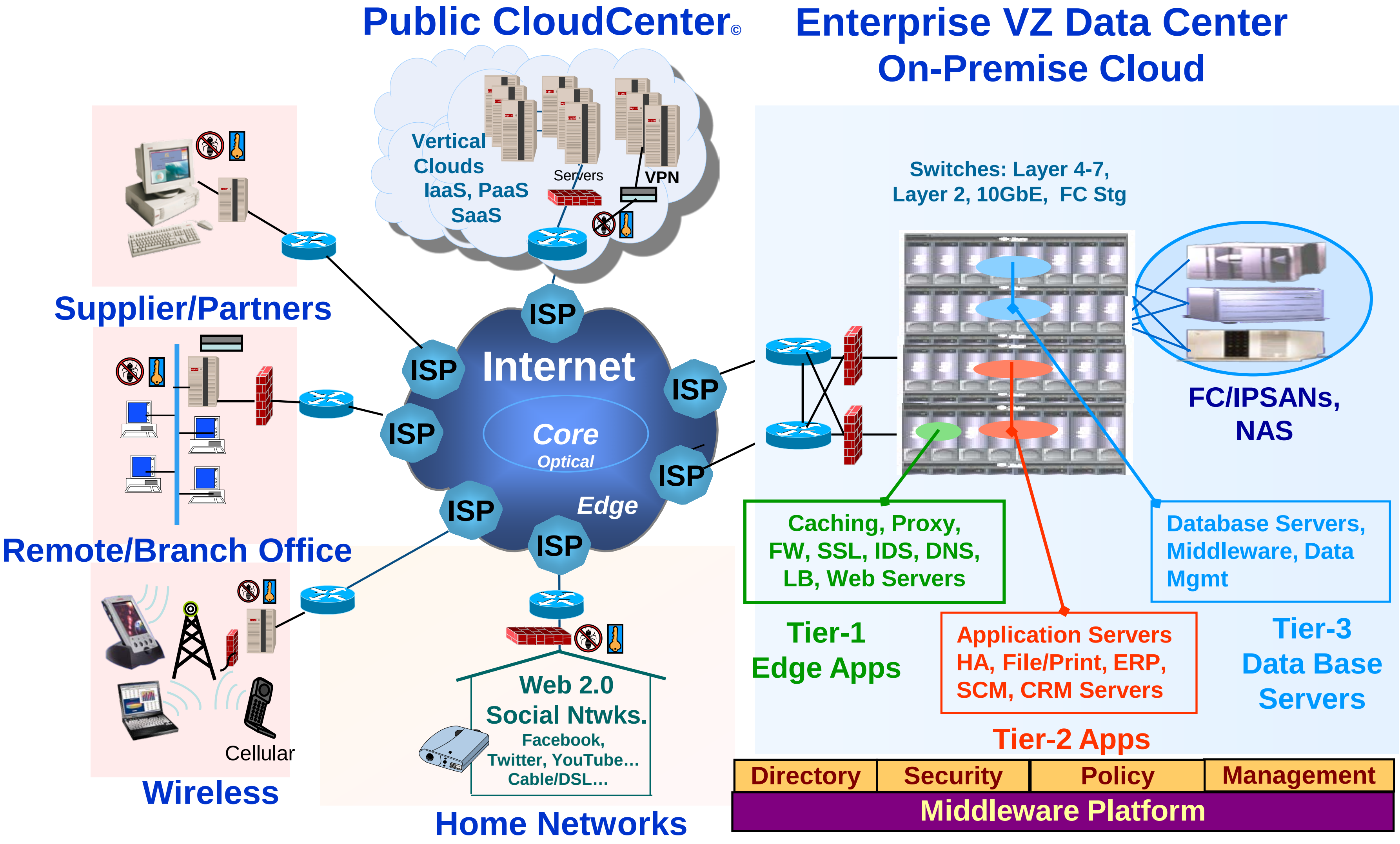
Cost, Availability, Performance, Scalability, Inter-operability, Manageability & Security

Standardization

Standard IT Infrastructure- Volume Economics HW/Syst SW

(Servers, Storage, Networking Devices, System Software (OS, MW & Data Mgmt. SW)

IT Infrastructure: DataCenters & Cloud



Public CloudCenter©

Enterprise VZ Data Center On-Premise Cloud

Vertical Clouds IaaS, PaaS SaaS

Servers

VPN

Switches: Layer 4-7, Layer 2, 10GbE, FC Stg

Supplier/Partners

Internet

Core Optical

Edge

FC/IPSANs, NAS

Remote/Branch Office

Caching, Proxy, FW, SSL, IDS, DNS, LB, Web Servers

Database Servers, Middleware, Data Mgmt

Tier-1 Edge Apps

Application Servers HA, File/Print, ERP, SCM, CRM Servers

Tier-3 Data Base Servers

Web 2.0 Social Ntwks. Facebook, Twitter, YouTube… Cable/DSL…

Cellular

Wireless

Tier-2 Apps

Directory	Security	Policy	Management

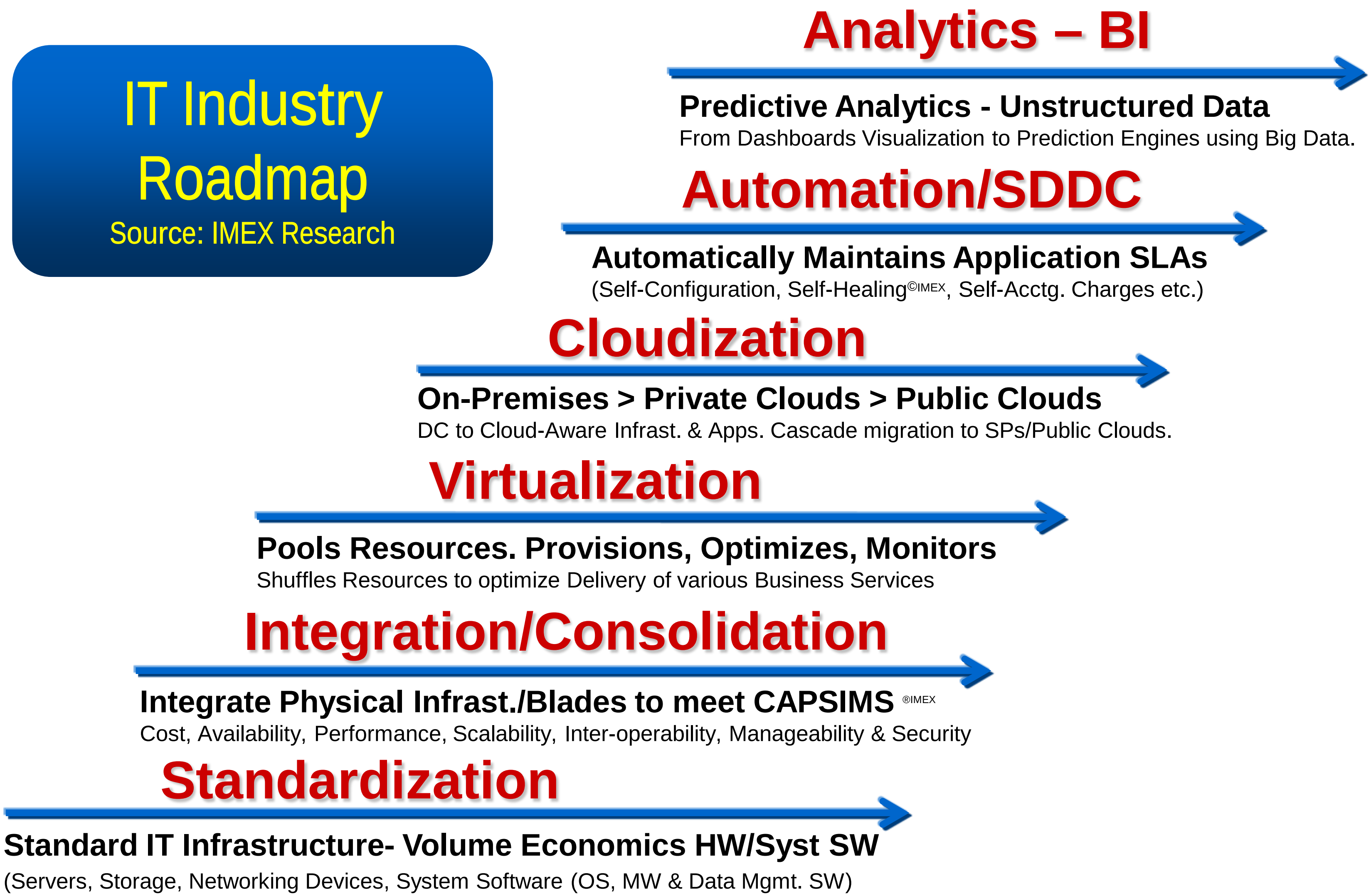
Middleware Platform

Home Networks

Source:: IMEX Research - Cloud Infrastructure Report ©2009-11

IT Industry Dynamics - Roadmap

IMEX RESEARCH.COM



IT Industry Roadmap

Source: IMEX Research

Analytics – BI

Predictive Analytics - Unstructured Data

From Dashboards Visualization to Prediction Engines using Big Data.

Automation/SDDC

Automatically Maintains Application SLAs

(Self-Configuration, Self-Healing©IMEX, Self-Acctg. Charges etc.)

Cloudization

On-Premises > Private Clouds > Public Clouds

DC to Cloud-Aware Infrast. & Apps. Cascade migration to SPs/Public Clouds.

Virtualization

Pools Resources. Provisions, Optimizes, Monitors

Shuffles Resources to optimize Delivery of various Business Services

Integration/Consolidation

Integrate Physical Infrast./Blades to meet CAPSIMS ®IMEX

Cost, Availability, Performance, Scalability, Inter-operability, Manageability & Security

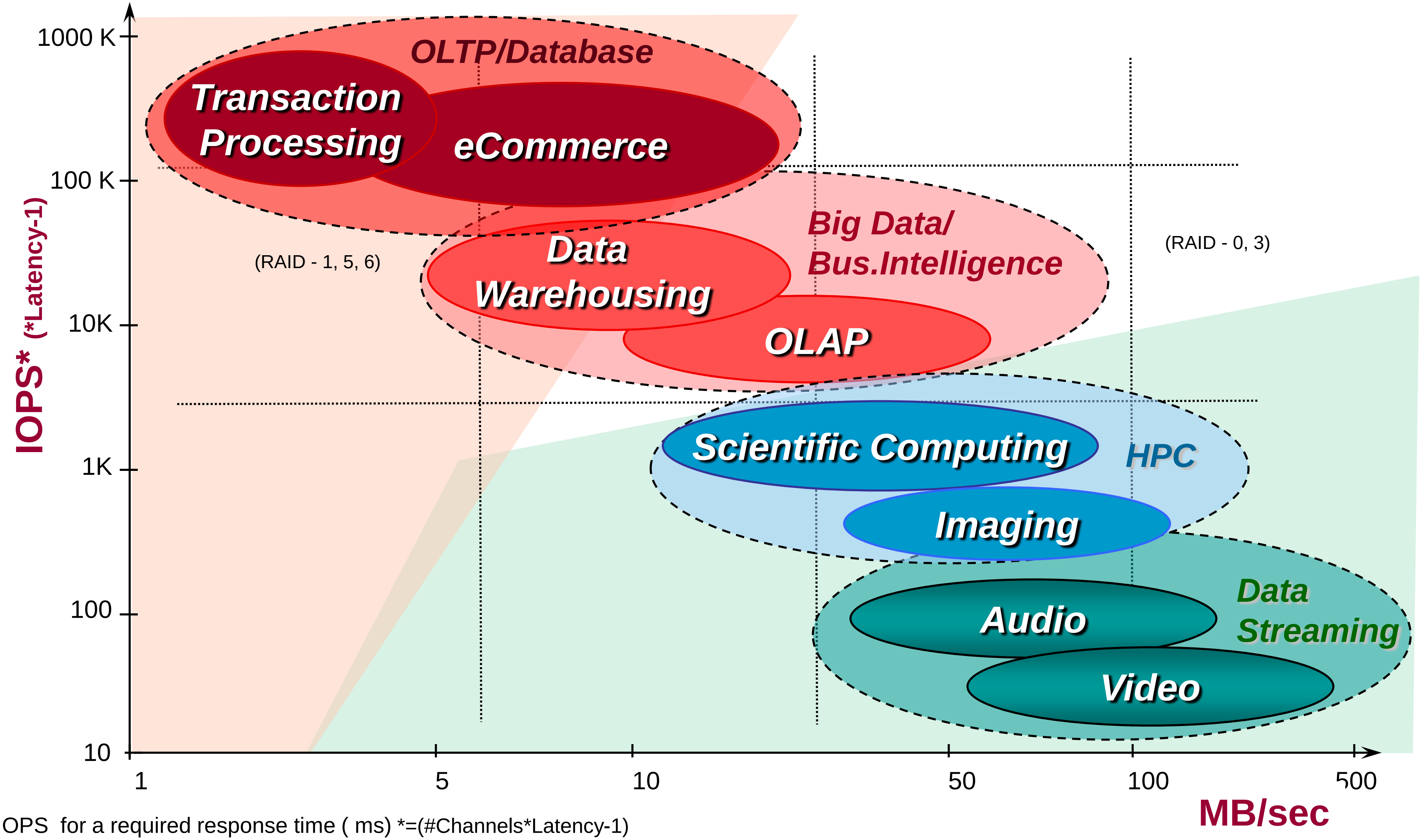
Standardization

Standard IT Infrastructure- Volume Economics HW/Syst SW

(Servers, Storage, Networking Devices, System Software (OS, MW & Data Mgmt. SW)

Workloads: Mapped on Infrastructure Metrics

IMEX RESEARCH.COM



*IOPS for a required response time (ms) *=(#Channels*Latency-1)

Workloads need Infrastructure Optimized for Cost, Availability, Performance …

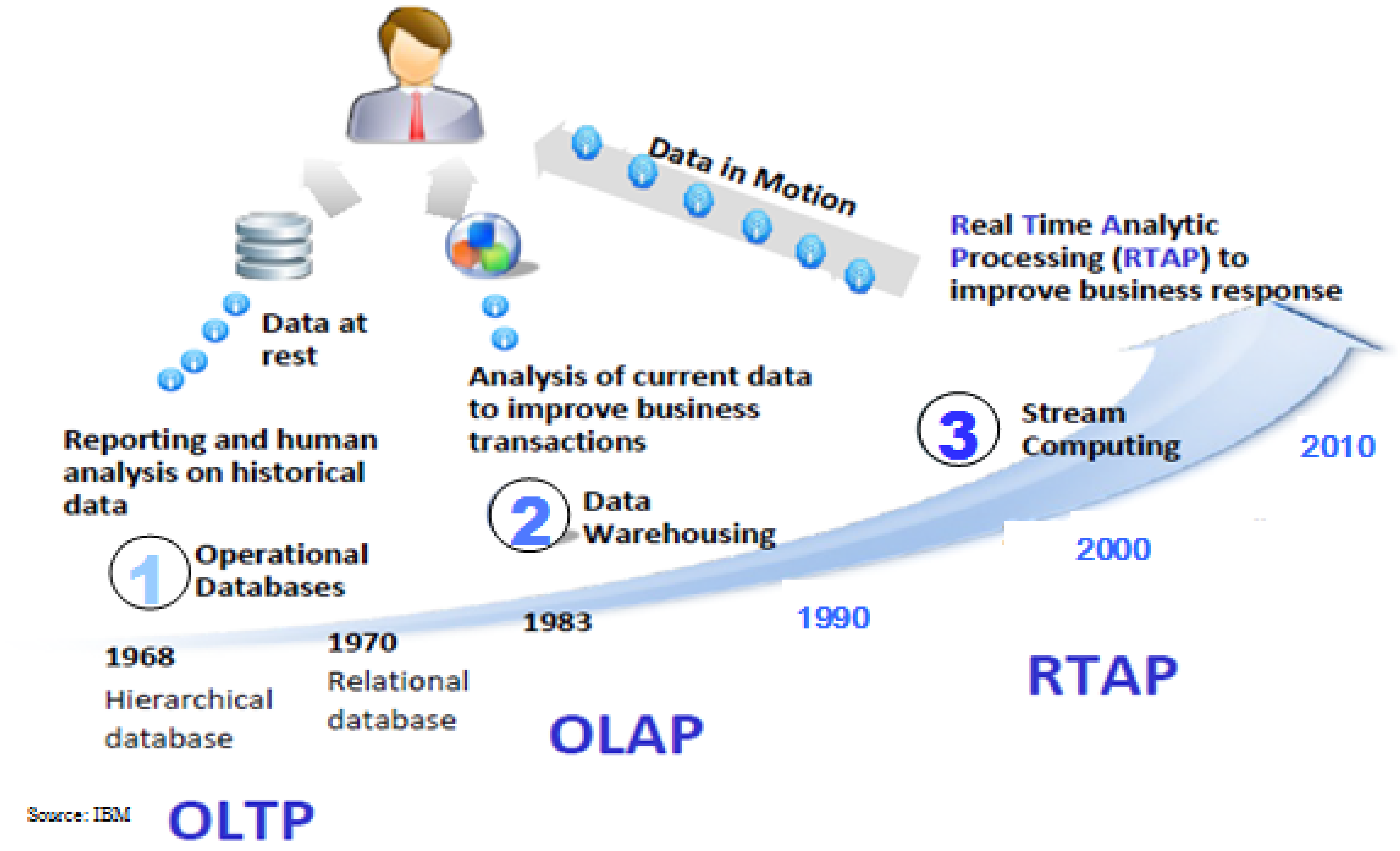
Workloads: I/O Characteristics

Storage performance, management and costs are big issues in running Databases

- **Data Warehousing Workloads are I/O intensive**
  - Predominantly read based with low hit ratios on buffer pools
  - High concurrent sequential and random read levels
    - ✓ Sequential Reads requires high I/O Bandwidth (MB/sec)
    - ✓ Random Reads require high IOPS
  - Write rates driven by life cycle management and sort operations
- **OLTP Workloads are strongly random I/O intensive**
  - Random I/O is more dominant
    - ✓ Read/write ratios of 80/20 are most common but can be 50/50
    - ✓ Difficult to build out test systems with sufficient I/O characteristics
- **Batch Workloads (Hadoop) are more write intensive**
  - Sequential Writes requires high I/O Bandwidth (MB/sec)
- **Backup & Recovery times are critical for these workloads**
  - Backup operations drive high level of sequential IO
  - Recovery operation drives high levels of random I/O

Source: IMEX Research SSD Industry Report ©2011

Driver : Need Real Time Analytics



Data in Motion

Real Time Analytic Processing (RTAP) to improve business response

Data at rest

Analysis of current data to improve business transactions

Reporting and human analysis on historical data

3 Stream Computing

2010

2 Data Warehousing

2000

1 Operational Databases

1990

1983

1970 Relational database

1968 Hierarchical database

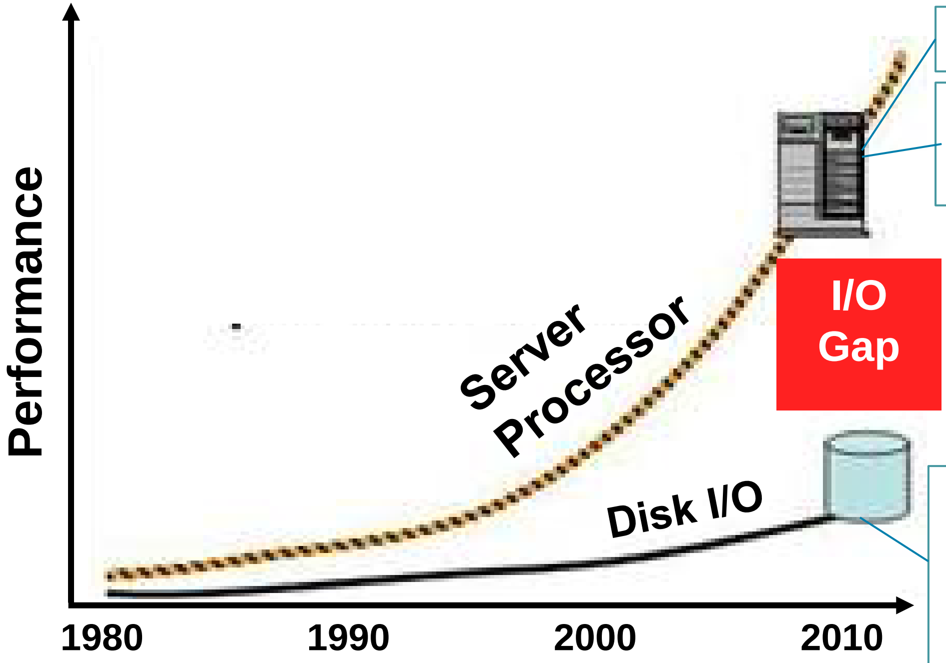
RTAP

OLAP

OLTP

Source: IBM

Issue: Server to Storage I/O Gap

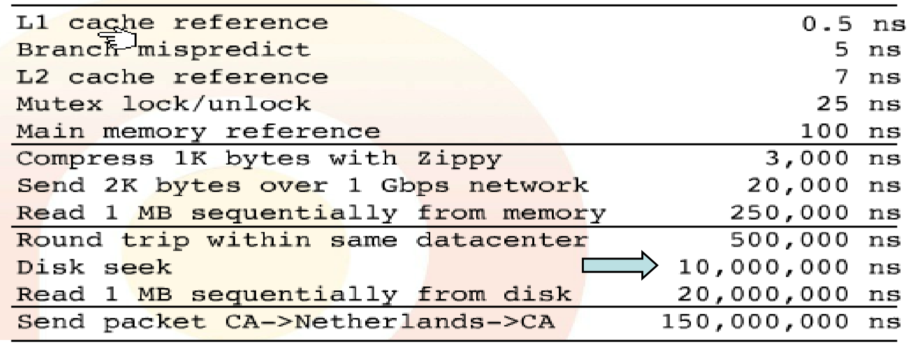


Memory - MultiSlots

Connect to Storage as Direct Attached Storage (Internal Storage)

For Each Disk Operation,
Millions of CPU Opns. or Thousands of Memory Opns. can be accomplished

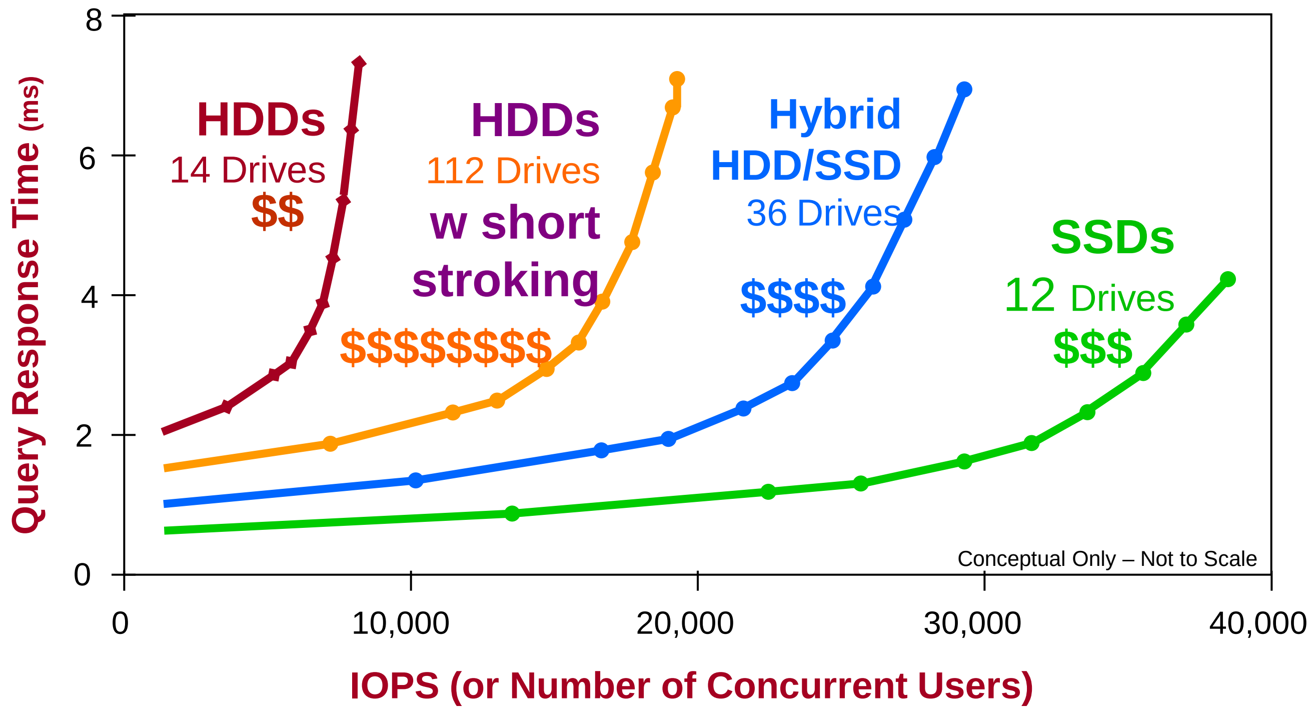
PCIe used for HBAs to connect to (External Shared Storage) via Storage Switches/Fabric as SAN/NAS on HDDs front ended by DRAM Cache creating a gap of 100,000x latency gap



L1 cache reference	0.5 ns
Branch mispredict	5 ns
L2 cache reference	7 ns
Mutex lock/unlock	25 ns
Main memory reference	100 ns
Compress 1K bytes with Zippy	3,000 ns
Send 2K bytes over 1 Gbps network	20,000 ns
Read 1 MB sequentially from memory	250,000 ns
Round trip within same datacenter	500,000 ns
Disk seek	10,000,000 ns
Read 1 MB sequentially from disk	20,000,000 ns
Send packet CA->Netherlands->CA	150,000,000 ns

A 7.2K/15k rpm HDD can do 100/140 IOPS

Solution: SSDs Improving DB Query Responses



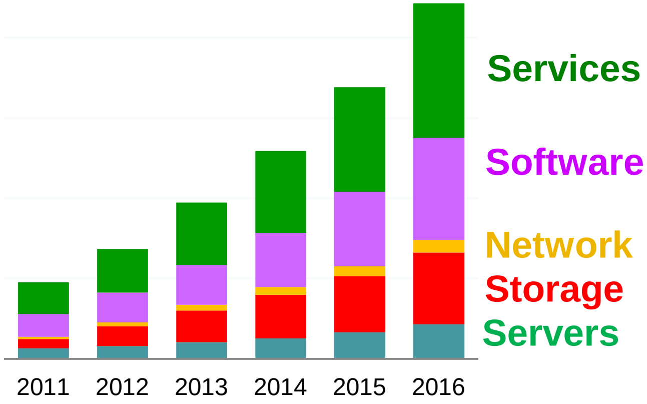
For a targeted query response time in DB & OLTP applications, many more concurrent users can be added cost-effectively when using SSDs or SSD + HDDs storage vs. adding more HDDs or short-stroking HDDs

Source: IMEX Research SSD Industry Report ©2011

Industry Trends: Impact on Infrastructure

IMEX RESEARCH.COM

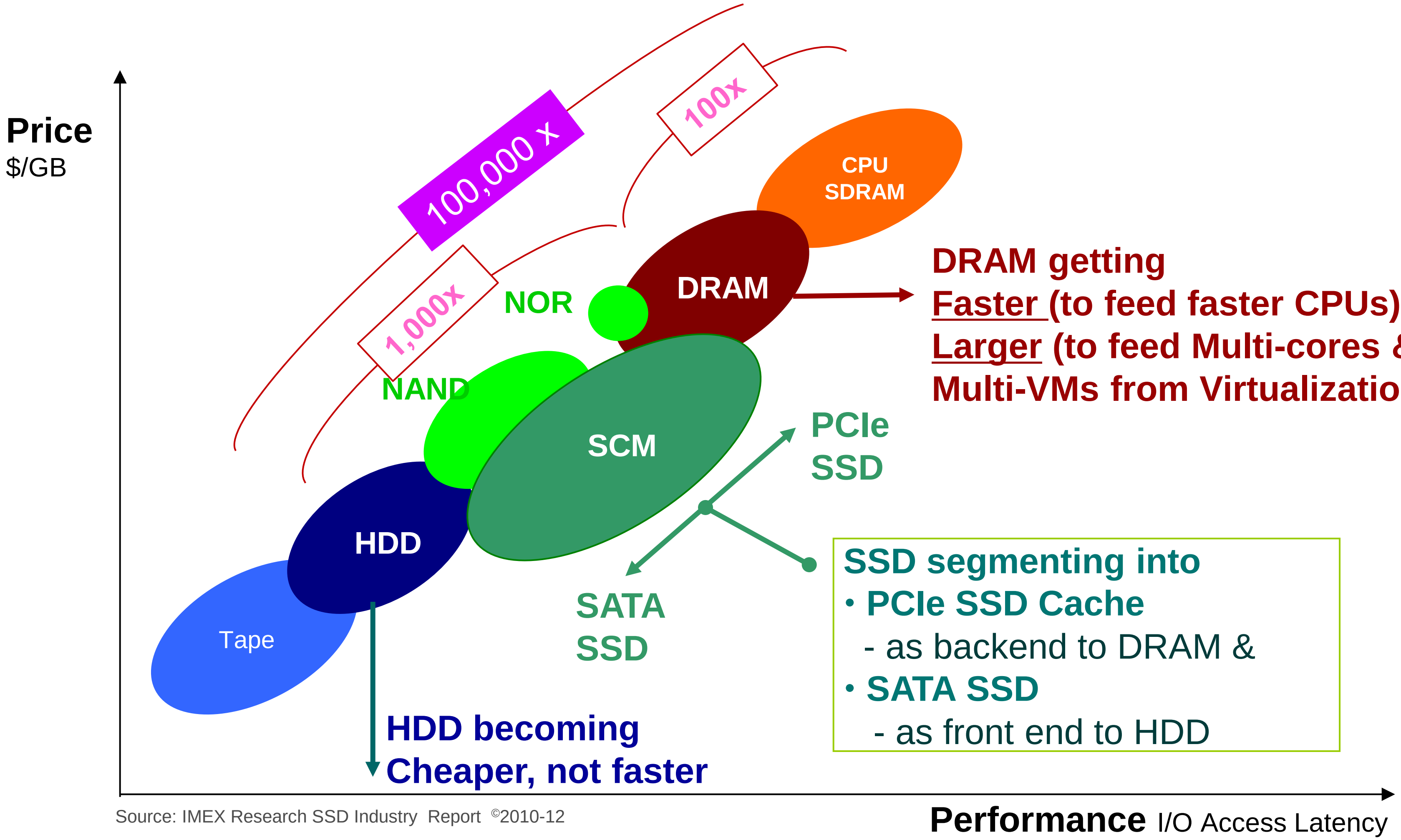
Systems & Services Market Revenues $B



Industry Dynamics	Impact on Infrastructure
Serviceable PCIe SSDs	Same Connector - PCIe+SAS/SATA
Lower Cost DRAM	NVMe based Flash Memory
BYOD / Boot Storms	Client Images on Servers
Big Data/RealTm Analytics	PCIe Servers based SSD
Server/Stg. Price/Perf.	New Storage Class Memory
Cloud Computing	New Protocols / REST, HTTP..
Multicores	Flash for Concurrent Multitasking
Power Efficiency	Green Memory
Virtualization	New Infrastrecture for Multi-VMs
Scale Out Clustering	Distributed Memory Architecture
Dense Blades	Fast, Low Power Memory
64 bit Computing	Larger Size Memories

Workloads need Infrastructure - Optimized for Cost, Availability, Performance …

Solution: SSDs Filling Price/Perf Gap



DRAM getting Faster (to feed faster CPUs) & Larger (to feed Multi-cores & Multi-VMs from Virtualization)

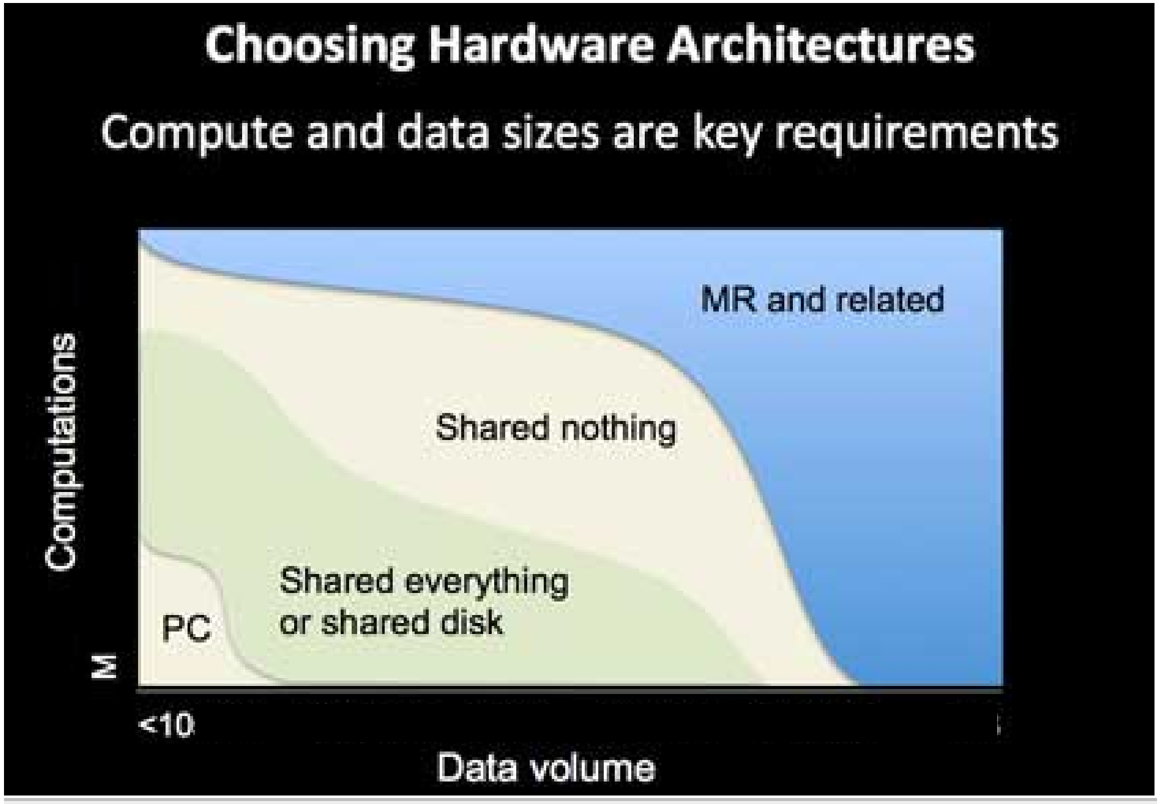
SSD segmenting into
- **PCIe SSD Cache**
  - as backend to DRAM &
- **SATA SSD**
  - as front end to HDD

HDD becoming Cheaper, not faster

Source: IMEX Research SSD Industry Report ©2010-12

Best Opportunity to fill the gap is for storage to be close to Server CPU.

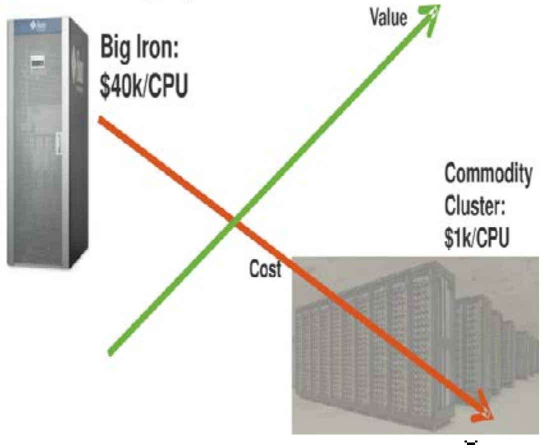
Innovations Roadmap: HW Technologies

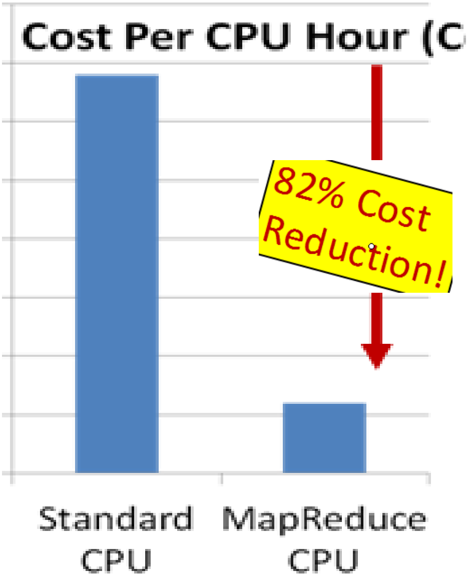


Choosing Hardware Architectures

Compute and data sizes are key requirements

- Hardware cost halving every 18mo



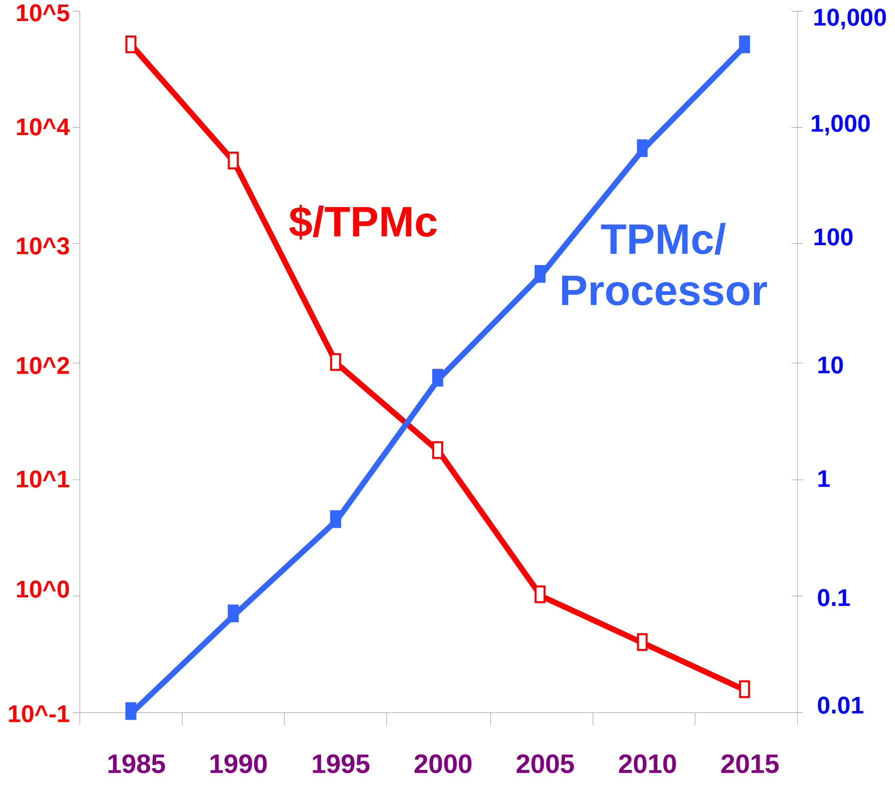


Cost Per CPU Hour (C

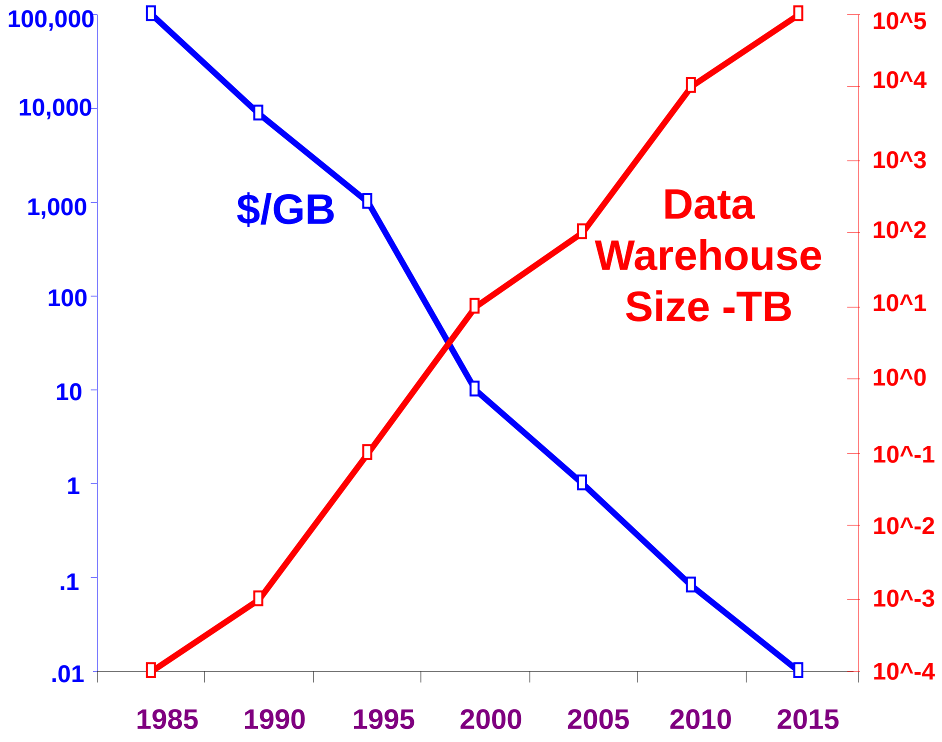
Innovations Roadmap – DB SW Technologies

IMEX RESEARCH.COM

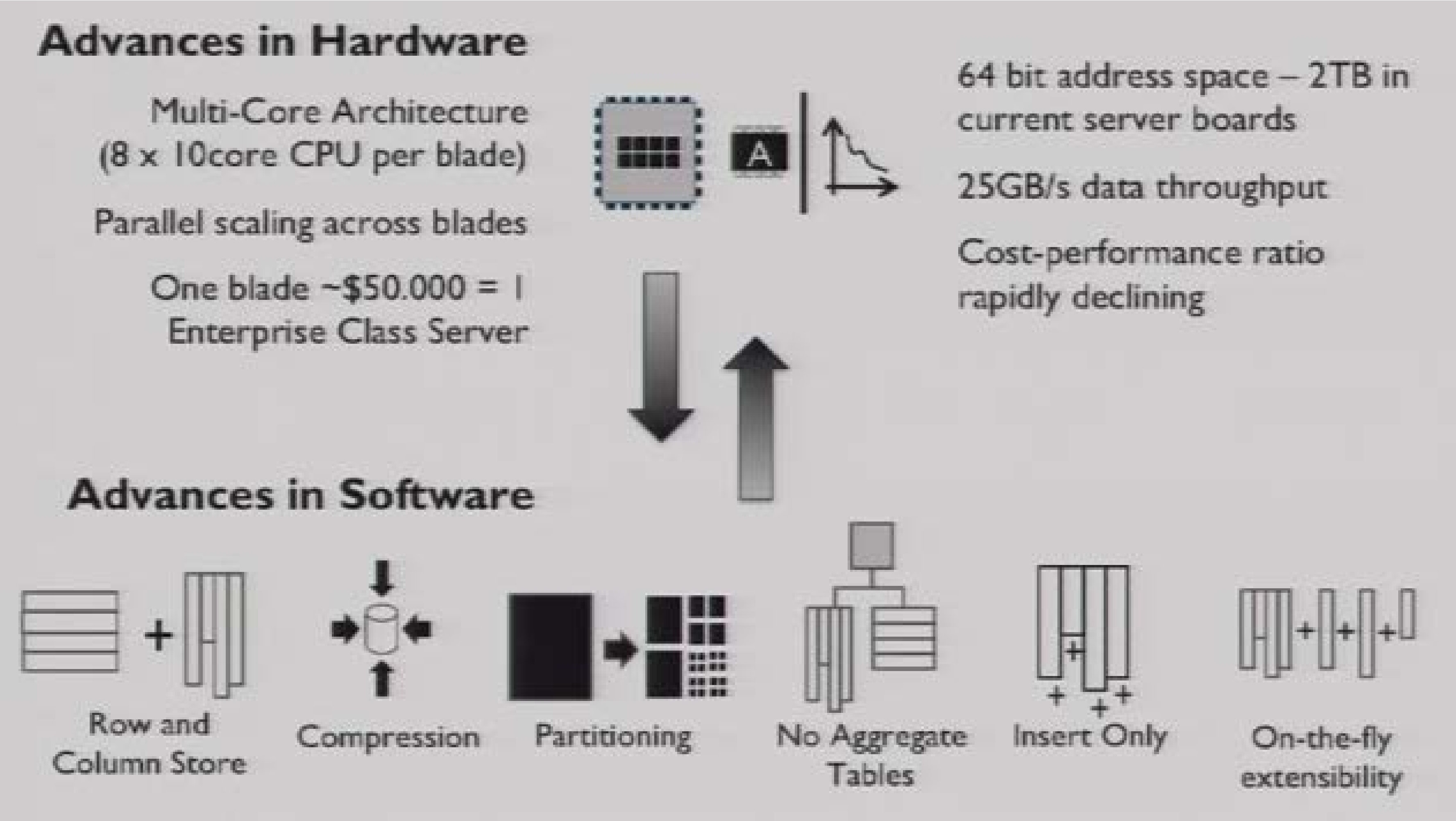
OLTP Database Innovation Progress



EDW/Big Data Database Innovation Progress



In-Memory Computing



Advances in Hardware

Multi-Core Architecture
(8 x 10core CPU per blade)

Parallel scaling across blades

One blade ~$50.000 = 1
Enterprise Class Server

64 bit address space – 2TB in
current server boards

25GB/s data throughput

Cost-performance ratio
rapidly declining

Advances in Software

Row and
Column Store

Compression

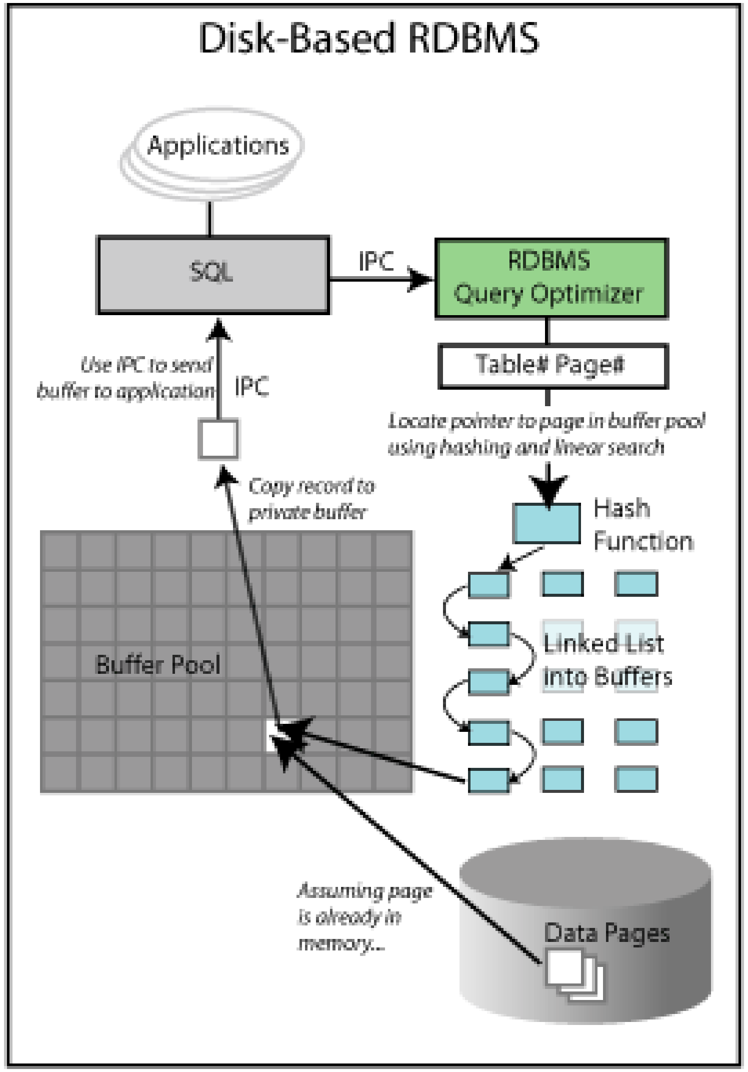
Partitioning

No Aggregate
Tables

Insert Only

On-the-fly
extensibility

Technology: Disk vs InMem DB Architecture



Disk-Based RDBMS

Applications

SQL

IPC

RDBMS Query Optimizer

Table# Page#

Locate pointer to page in buffer pool using hashing and linear search

Hash Function

Linked List Into Buffers

Use IPC to send buffer to application

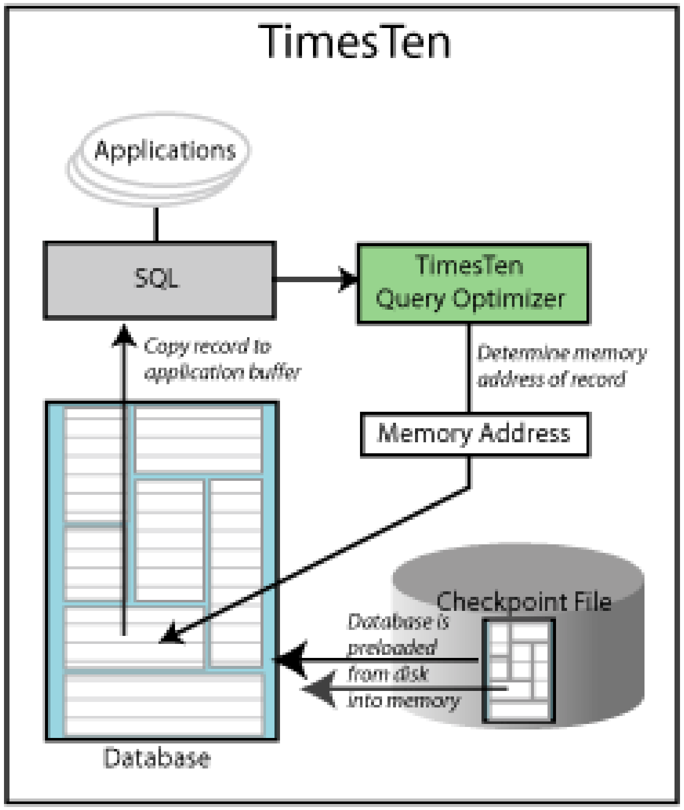
IPC

Copy record to private buffer

Buffer Pool

Assuming page is already in memory...

Data Pages



TimesTen

Applications

SQL

TimesTen Query Optimizer

Copy record to application buffer

Determine memory address of record

Memory Address

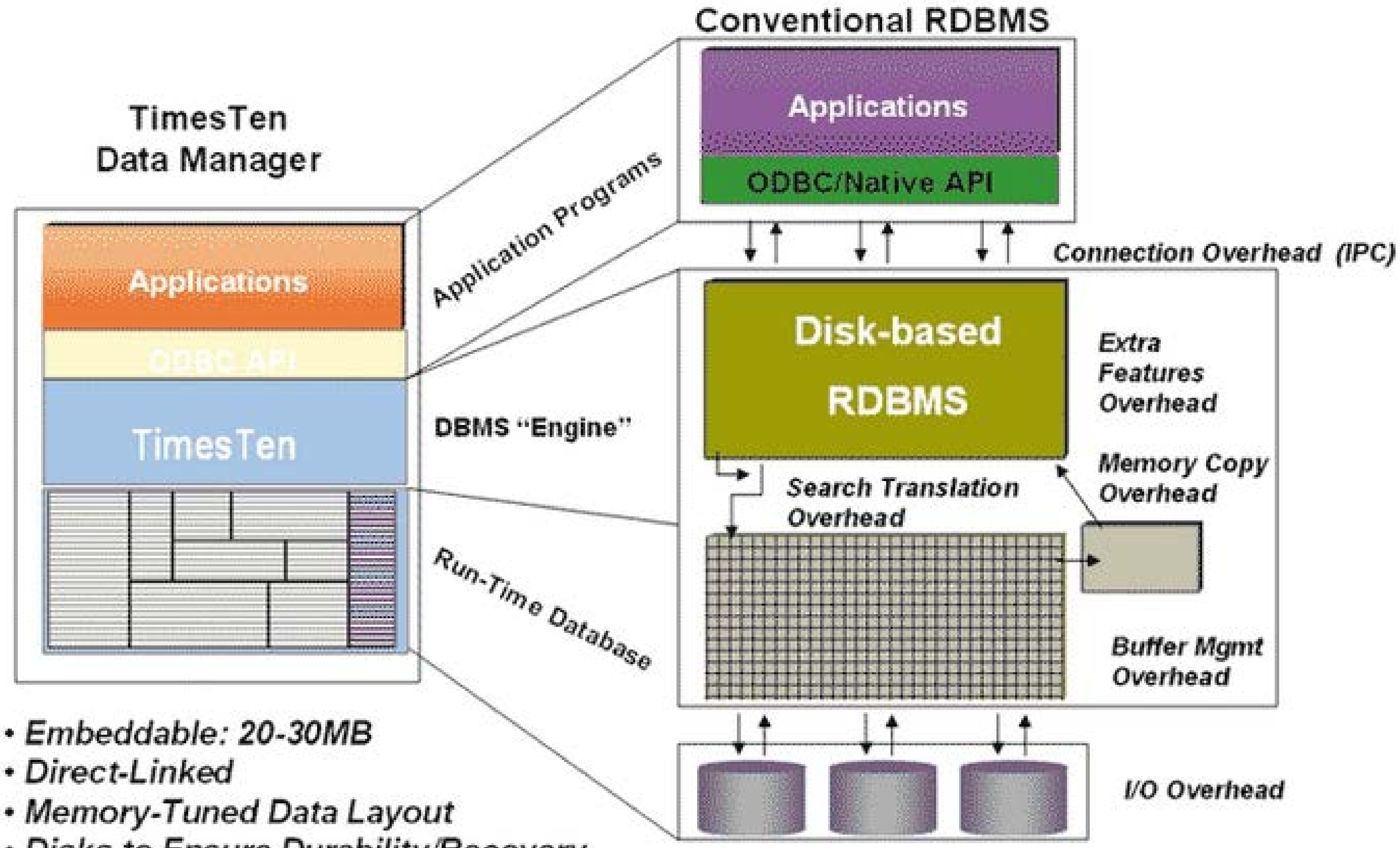
Checkpoint File

Database is preloaded from disk into memory

Database

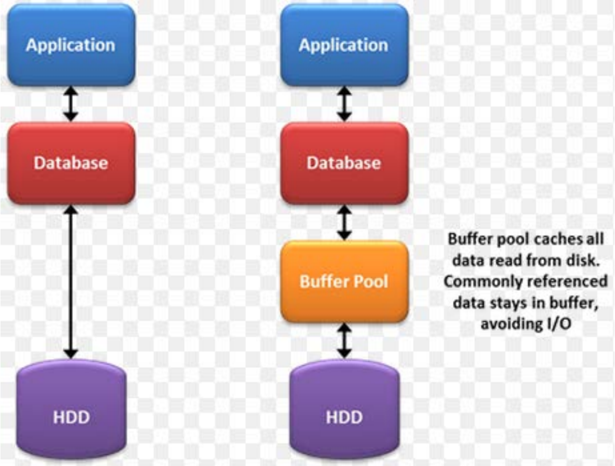
Technology: RDBMS vs. In-Memory DBMS

IMEX RESEARCH.COM

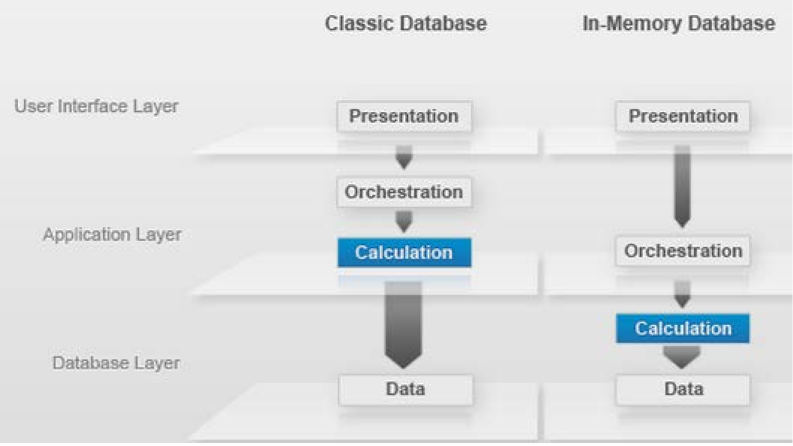


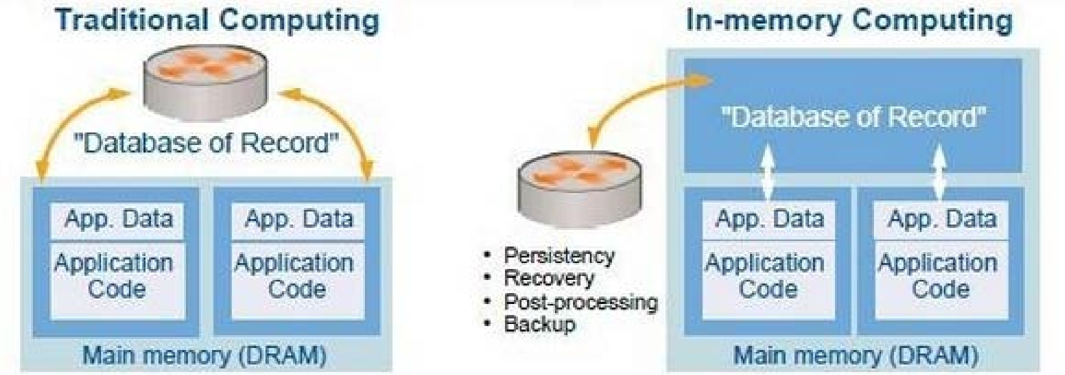
- *Embeddable: 20-30MB*
- *Direct-Linked*
- *Memory-Tuned Data Layout*
- *Disks to Ensure Durability/Recovery*

Technology:Legacy vs In-Memory Computing



Buffer pool caches all data read from disk. Commonly referenced data stays in buffer, avoiding I/O

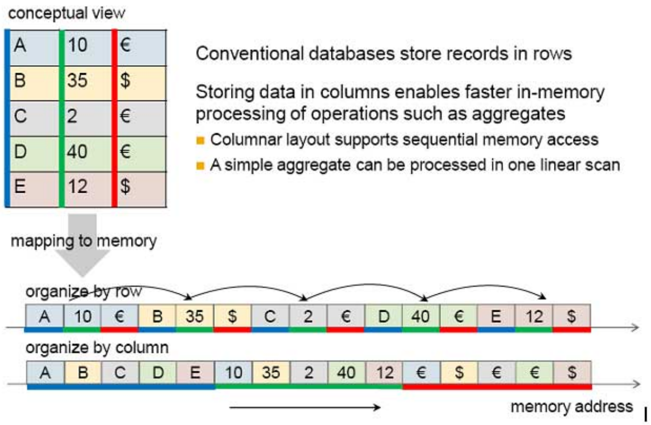




- Persistency
- Recovery
- Post-processing
- Backup

Why Now?

- 64-bit processors can address **up to 16 exabytes of data**
- DRAM production costs **drop by 32% every 12 months**
- 1GB of NAND flash memory **average price is 56$ cents***
- Commodity hardware provide **multi terabyte of DRAM**
- In-memory-enabling **software is available and proven**
- IMC software is often **embedded in products/services**


conceptual view

A	10	€
B	35	$
C	2	€
D	40	€
E	12	$

Conventional databases store records in rows

Storing data in columns enables faster in-memory processing of operations such as aggregates

- Columnar layout supports sequential memory access
- A simple aggregate can be processed in one linear scan

mapping to memory

organize by row

A	10	€	B	35	$	C	2	€	D	40	€	E	12	$

organize by column

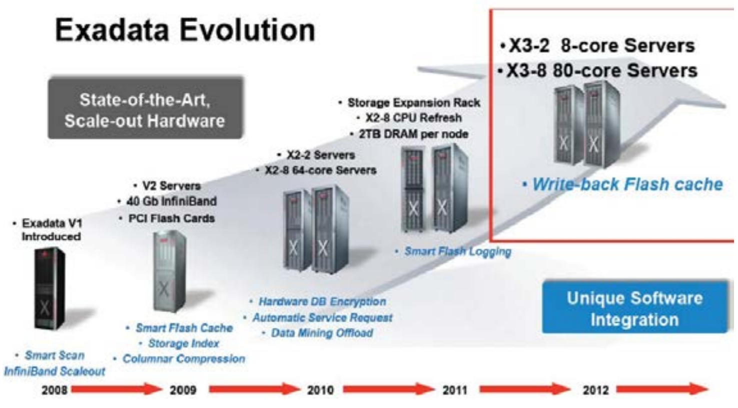
A	B	C	D	E	10	35	2	40	12	€	$	€	€	$

memory address

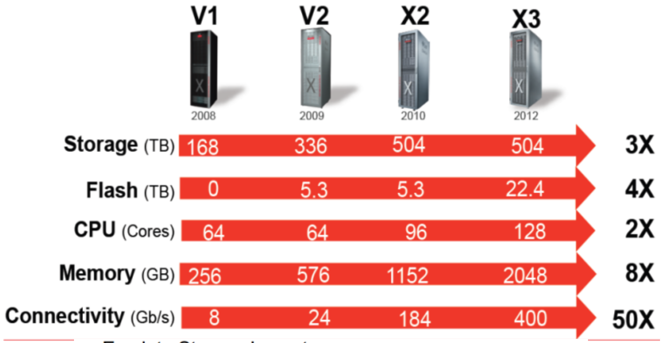
Oracle vs SAP vs IBM DB

IMEX RESEARCH.COM

Exadata Evolution

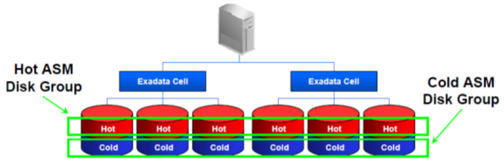


Hardware Generation Advances



	V1 2008	V2 2009	X2 2010	X3 2012	
Storage (TB)	168	336	504	504	3X
Flash (TB)	0	5.3	5.3	22.4	4X
CPU (Cores)	64	64	96	128	2X
Memory (GB)	256	576	1152	2048	8X
Connectivity (Gb/s)	8	24	184	400	50X

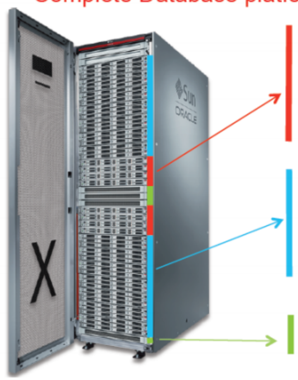
Exadata Storage Layout



- Two ASM disk groups defined
  - One for the active, or "hot" portion, of the database and a second for the "cold" or inactive portion
- ASM striping evenly distributes I/O across the disk group
- ASM mirroring is used protect against disk failures
  - Optional for one or both disk groups

Exadata Architecture

Complete Database platform using standard servers for Compute and Storage

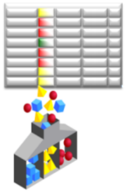


Scale-Out Database Servers
- 2-socket or 8-socket Xeon database servers
- Oracle Database, ASM, RAC; Linux or Solaris
- Standard Ethernet to data center

Scale-Out Intelligent Storage Servers
- 2-socket storage servers, Exadata Storage Software
- Up to 500 terabytes disk per rack
- 56 PCI Flash memory cards per rack

InfiniBand Network
- Unified internal connectivity (40 Gb/sec)

Exadata Hybrid Columnar Compression
Highest Capacity, Lowest Cost



- Data is organized and cc
  - Dramatically better com
- Speed Optimized **Query** Warehousing
  - 10X compression typic
  - Runs faster because o
- Space Optimized **Archiv** infrequently accessed da
  - 15X to 50X compressic

Faster and Simpler
Backup, DR, Caching, Reorg, Clone

Benefits Multiply

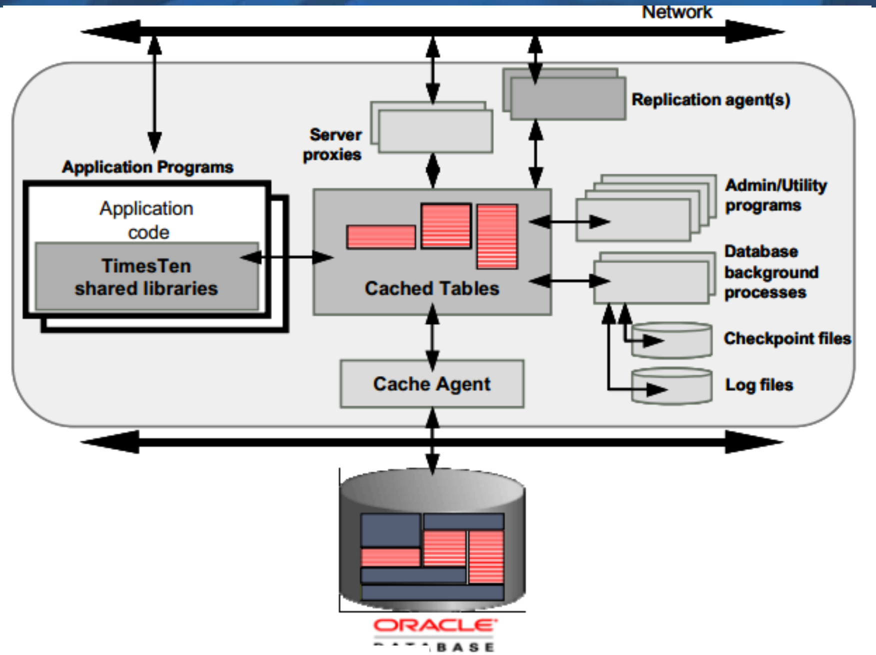
Exadata Smart Flash Cache
Extreme Performance OLTP

- Exadata has 5 TB / 22.3 TB of flash
  - 56 Flash PCI cards avoid disk controller bottlenecks
- Intelligently manages flash
  - Smart Flash Cache holds hot data
  - Gives speed of flash, cost of disk
- Exadata flash cache achieves:
  - Over 1 million IO/sec from SQL (8K)
  - Sub-millisecond response times
  - 50 GB/sec query throughput

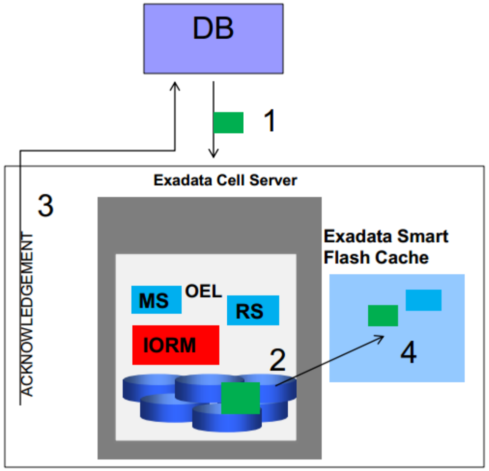
Exadata Summary

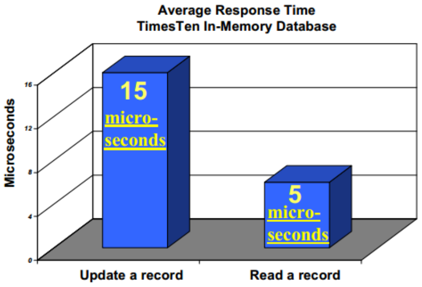
- Best for OLTP
  - Smart Flash Cache
  - 1 Million I/Os per Second
- Best for Warehousing
  - Intelligent Scale-Out storage
    - 10x faster queries
  - 10x Data Compression
- Best for Consolidation
  - Terabytes of Memory
  - Mix OLTP, DW, batch, reporting in single machine

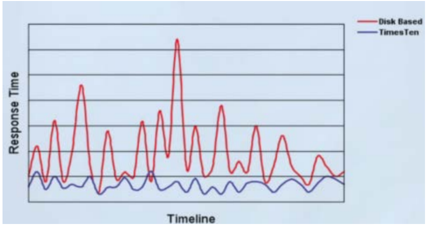
Competition: Oracle DB Architecture

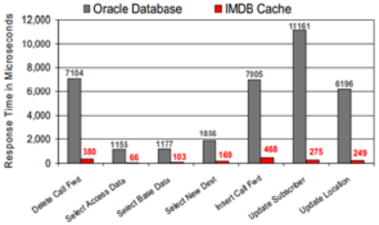


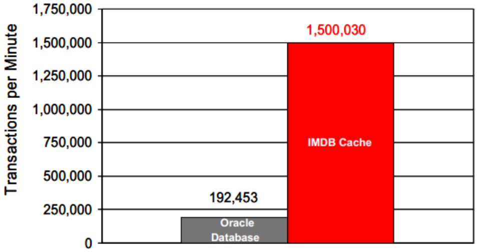
Exadata Smart Flash Cache – Write Operation



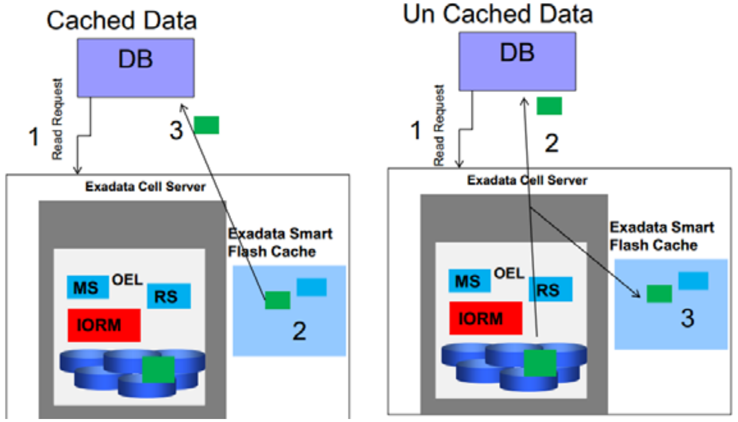








Exadata Smart Flash Cache – Read



Competition: SAP/HANA (Multi-Applications)

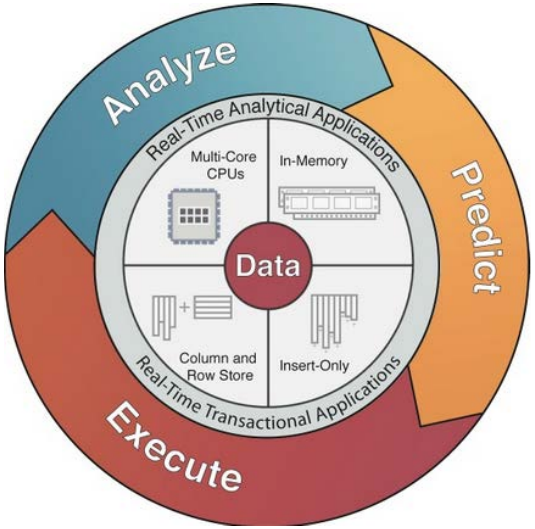
A Converged DB System

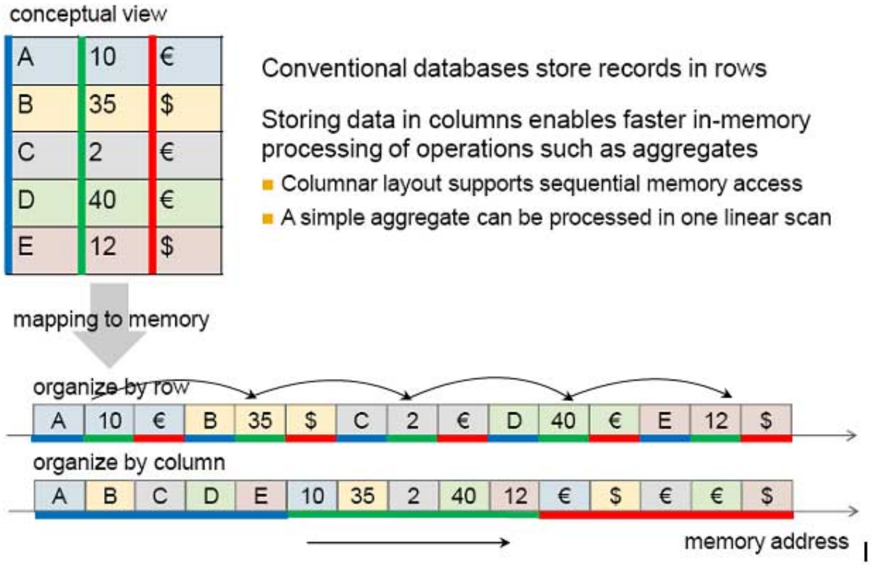
- **In-memory database combining transactional data processing, analytical data processing, and application logic processing functionality in memory**.
- **A full DBMS with a standard SQL interface, high availability, transactional isolation and recovery (ACID properties)**
- **both row-based and column-based stores within the same engine** (row-based storage is good for transactional applications, while column-based storage is better for reports and analytics. Column-based storage compresses the better too.)
- **massively parallel execution using multicore processors**, SAP HANA optimizes the SQL which scales well with the number of cores. Aggregation operations by spawning a number of threads that act in parallel, each of which has equal access to the data resident on the memory on that node
- Additional functions - **freestyle search** (as SQL extensions). BI applications using MDX for Microsoft Excel & Consumer Services plus internal I/F for BusinessObjects
- **prepackaged algorithms in the predictive analysis library** of SAP HANA to perform advanced statistical calculations
- **built-in text support**, from its predecessor BI Accelerator that was based on the TREX search engine and Inxight functionality integrated into HANA text functions.

Competition: SAP/HANA (Multi-Applications)

- supports distribution across hosts, where large tables may be partitioned to be processed in parallel. DB “engine” of the SAP HANA Analytics appliance as well
- HANA’s combination of a row and column store is fundamentally different from any other database engine on the market today, which allows it to perform OLTP and analytics processing in memory, at the same time.
- Avoids CPU waiting info from Memory through its unique CPU-cache-aware algorithms and data structures that there is as much useful data in the CPU caches as possible,.
- it uses late materialization to decompress columnar structures as late as possible, or to run operations directly on the compressed data
- also sold as an appliance on Intel Xeon CPUs leveraging insights into Intel’s HyperThreading, Turbo Boost and Threading Building Blocks
- High Performance Analytic Appliance can perform large-scale data analyses on 500 billion records in less than a minute, taking analytics to an entirely new dimension
- represents a complete data warehouse in RAM, and as a result, much accelerated real-time analytics.
- .

Technology: In-Memory Computing




conceptual view

A	10	€
B	35	$
C	2	€
D	40	€
E	12	$

Conventional databases store records in rows

Storing data in columns enables faster in-memory processing of operations such as aggregates

- Columnar layout supports sequential memory access
- A simple aggregate can be processed in one linear scan

mapping to memory

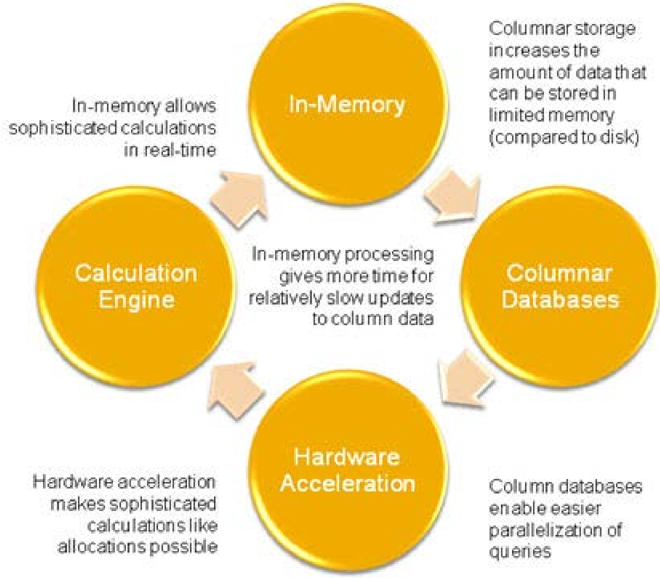
organize by row

A	10	€	B	35	$	C	2	€	D	40	€	E	12	$

organize by column

A	B	C	D	E	10	35	2	40	12	€	$	€	€	$

memory address



Columnar storage increases the amount of data that can be stored in limited memory (compared to disk)

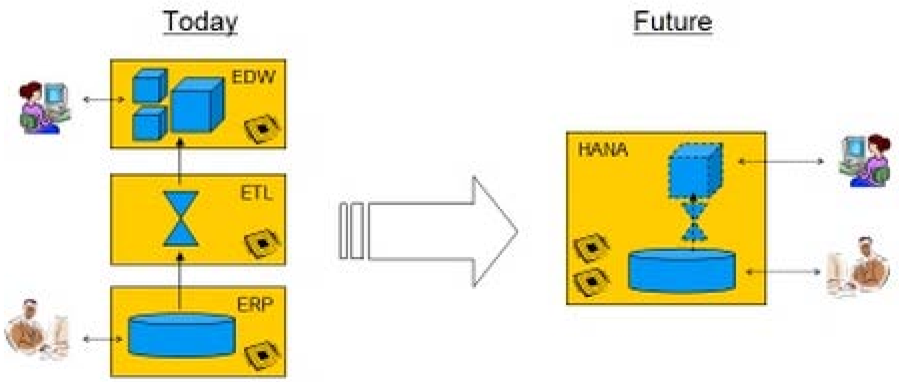
In-memory allows sophisticated calculations in real-time

In-memory processing gives more time for relatively slow updates to column data

Hardware acceleration makes sophisticated calculations like allocations possible

Column databases enable easier parallelization of queries

single HANA platform for all business processing



Trends: In-Memory Computing Adoption

SUSE HANA Certified Hardware

Pre-load SUSE Linux Enterprise Server for SAP Applications

IBM hp

FUJITSU DELL

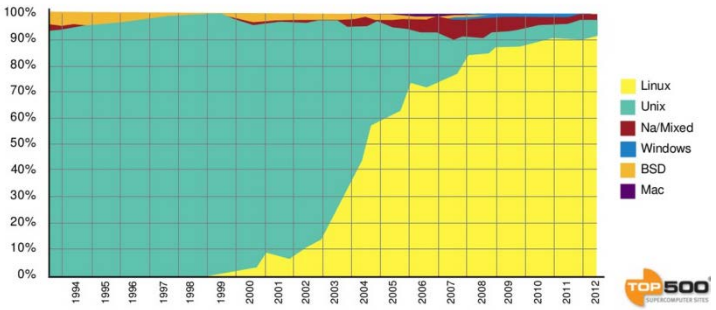
CISCO HITACHI

HUAWEI NEC

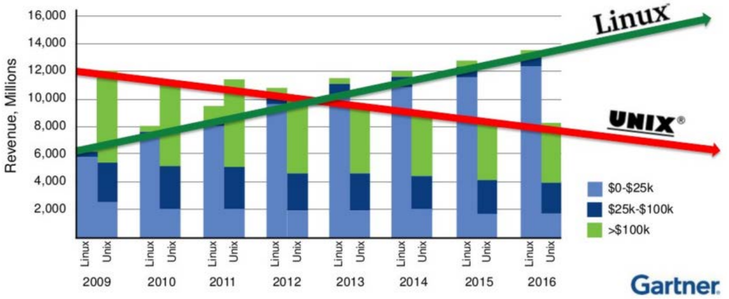
1&1 amazon web services FUJITSU

IBM Windows Azure

OS used on Top 500 Supercomputers

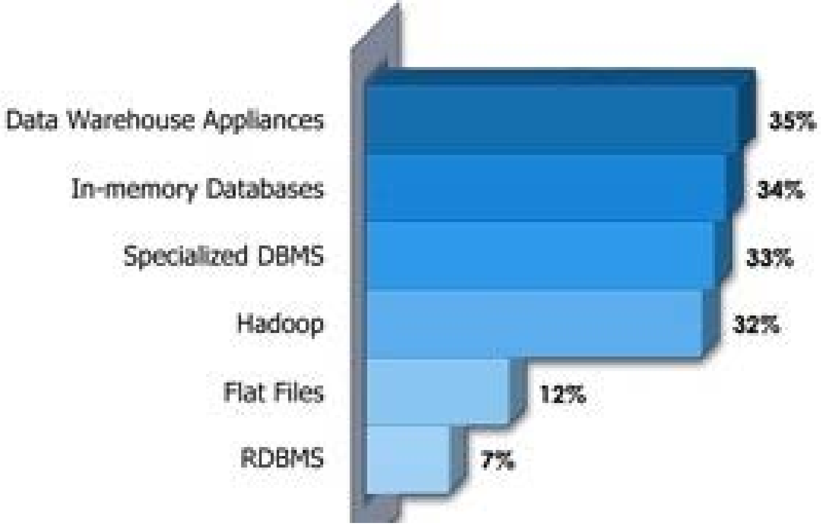


Unix Vs. Linux Trend Lines



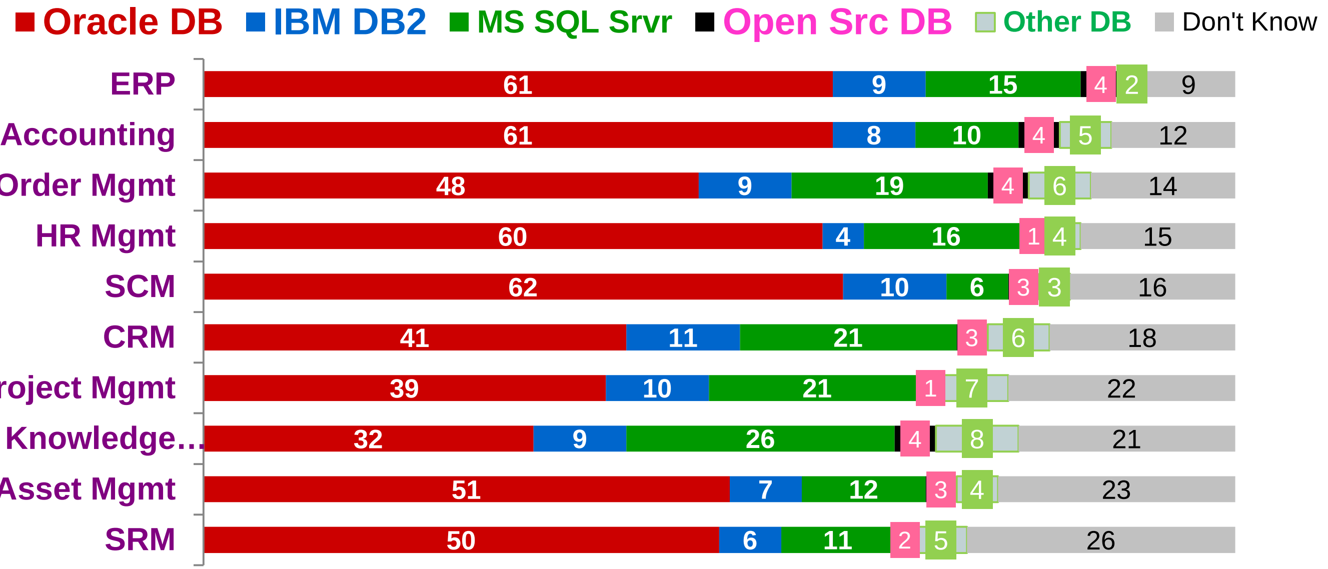
Big Data Technologies Planned

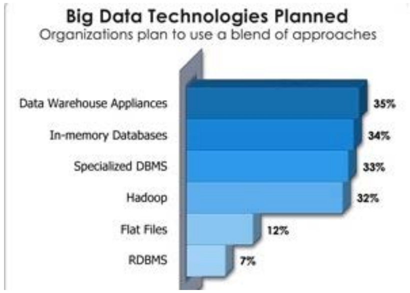
Organizations plan to use a blend of approaches



Trends: In-Memory DB Computing

Primary DB for Each Application





System Architecture
for
In-Memory Database

Anil Vasudeva
President & Chief Analyst
imex@imexresearch.com
408-268-0800

IMEX
RESEARCH.COM