



# **Driving Innovation Through the Information Infrastructure**

**SPRING 2011**

# **User Case Study: How Much Information? “Big Data” Research Findings for Enterprise and Consumer Information**

James Short, Roger Bohn and Chaitan Baru

UC San Diego, San Diego Supercomputer Center

Thursday, April 7 11:15 - Noon



# What is HMI?

- A research program at UCSD and SDSC
- Our goal: create a census of the World's information
  - How much is there?
  - Of what types?
  - How is it created and where does it go?
- Measuring data and information
  - Inexact science
  - Assumptions and methods key

# Our Agenda This Morning

- Review HMI's Greatest Hits
  - Consumer Information Report
  - (new) Enterprise Server Report
- Including:
  - A new way to measure server activity that aggregates capacity over very different kinds of servers, using a common metric
  - New data on price/performance across different benchmarks and server classes
  - Review and discuss alternative views of server load types and load factors
  - Identify methods for defining and measuring data and information value
- Ongoing and Future Research



# Consumers





facebook

Facebookで友達や家族と近況を知らせ合ったり、メッセージを交換したりして、楽しく交流しましょう。

[Digital Scholarshipさんが書いたノート](#)

brisbanetimes

Feeling bombarded with information? You're not imagining it  
BENNY EVANGELISTA OF THE SAN FRANCISCO CHRONICLE  
December 10, 2009



カリフォルニア大学サンディエゴ校 (UCSD)  
2008年に米国人がコンピュータ・テレビ・ラ  
ゲームなどにより一日に接した(消費した)  
Information? 2009 Report on American Co  
イト



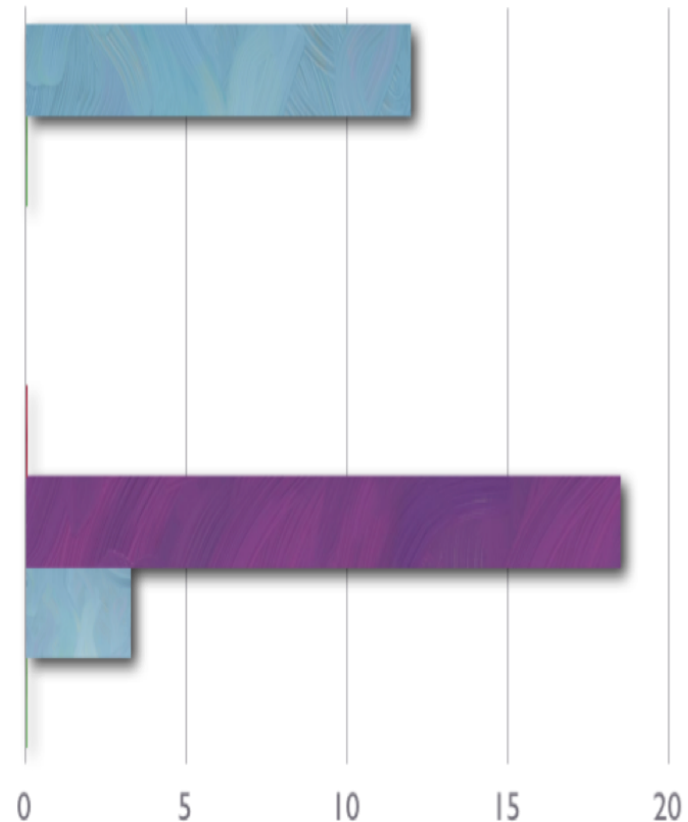
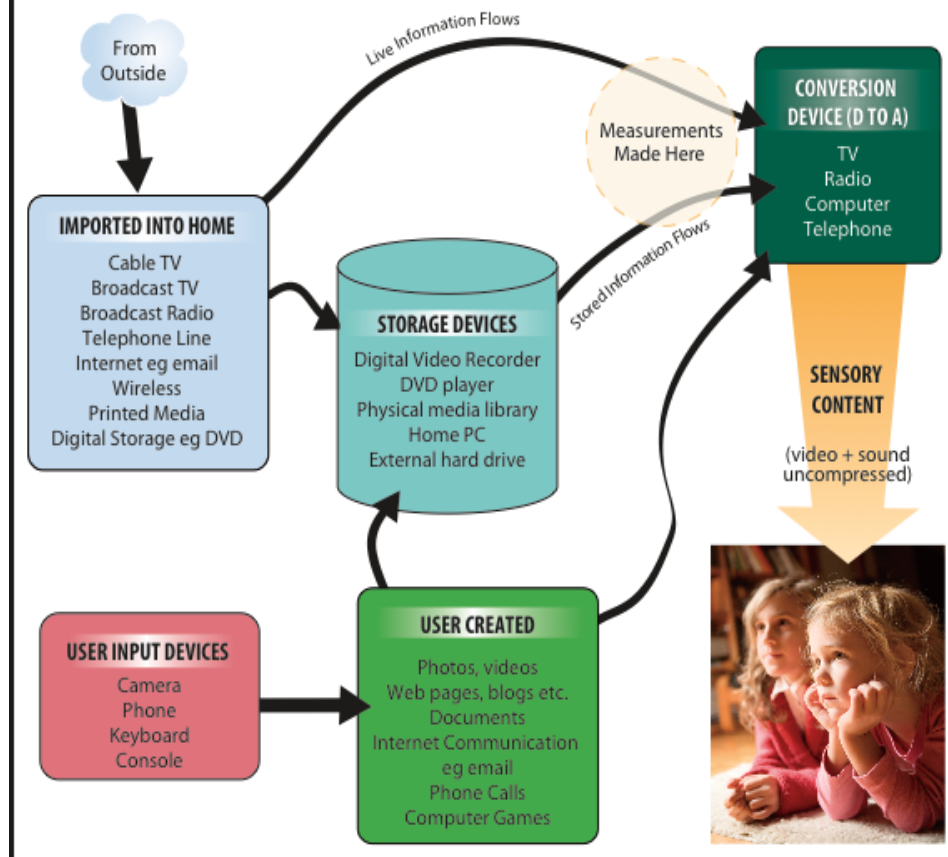
**U.S. households consumed 3.6 zettabytes of information in 2008**

**Dec. 10, 2009 (1:23 pm) By: [Brian Osborne](#)**

# Information to American consumers: 2008 average day

■ All TV ■ Radio ■ Phone ■ Print ■ Computer ■ Comp. game ■ Movies ■ Recorded music

**Figure 1: Information Flows In A Home**





# 34 GB, 12 hours, 100000 words

	Infoc in GB/day	Infow in words/day	Hours per day
TV (incl. DVR, Internet, mobile)	12.0	44,342	4.9
Radio	0.1	8,315	2.2
Phone	0.01	5,269	0.7
Print	0.01	8,659	0.6
Computer	0.08	27,122	1.9
Computer games	18.5	2,459	0.9
Movies	3.3	198	0.03
Recorded music	0.08	1,112	0.45
Total	34.0	~100,000	~ 12





# Slow Growth: 5% CAGR

	1960 to 1980	1980 to 2008
INFO <sub>Hours</sub>	3.9%	2.6%
INFO <sub>Words</sub>	3.7%	3.0%
INFO <sub>c</sub> (bytes)	2.9%	<b>5.4%</b>
US Population	1.1%	1.0%
GDP \$/capita	3.6%	2.9%





# Companies





# Summary: How Many Bytes?

**Table 1 World Server Information**

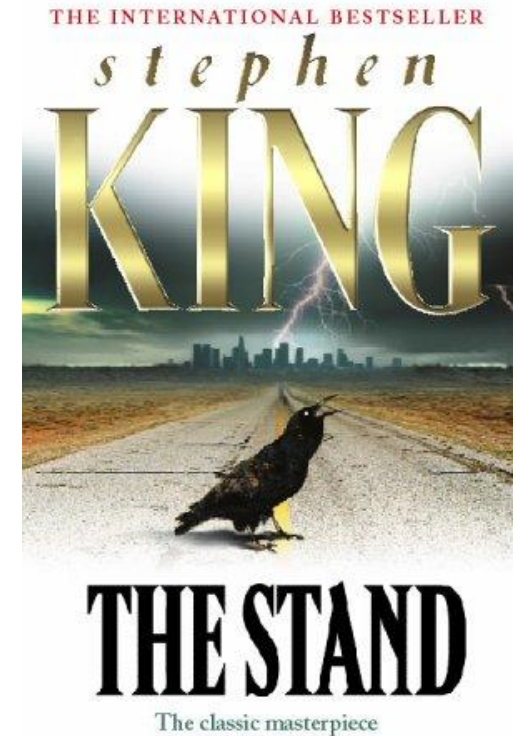
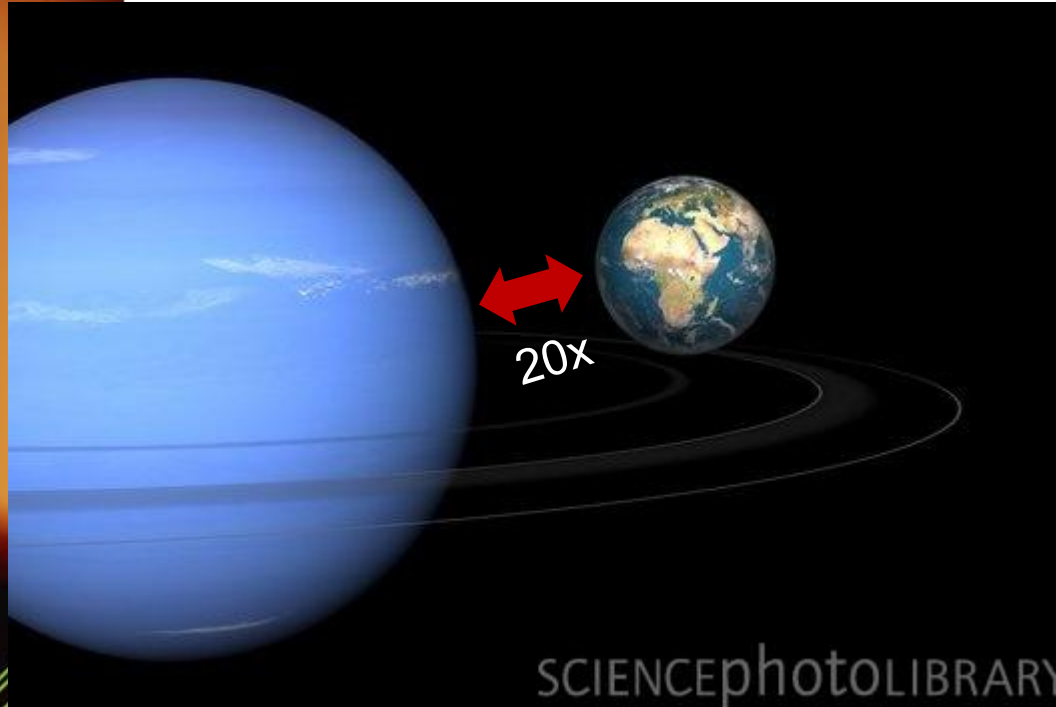
What is measured	World 2008 Total	Notes
Bytes processed plus delivered	9.57 zettabytes	
Bytes per worker per year per day	3.01 terabytes 12.0 gigabytes	3.18 billion workers in world labor force
Bytes per company per year	63.4 terabytes	151 million world businesses registered

**Table 3: World Server Information by  
Server Class 2008**

	Entry-level	Midrange	High-end	Total
<b>Total Bytes by Server Class</b> (in zettabytes)	<b>6.31</b>	<b>2.80</b>	<b>.451</b>	<b>9.57</b>



# How Much is 9.57 Zettabytes?



- 2.5 megabytes in Stephen King's Longest Novel
- Would need to stack novels from here to Neptune 20x to equal one year of server information

# Enterprise Information Could Include:

~~Data delivered to workers — screen-based~~

~~Data stored on storage media (what about redundancy?)~~

Data used (inputs & outputs - to inform, to process something, to take action)

~~Data in embedded processors in office & industrial machines~~

## Our model:

“Information” = **Data** processed plus delivered for use

Measured as: **Bytes** processed plus delivered by servers

Servers = **World** installed base in 2008



# Measuring “Work”

**WORK = WORK  
Performed  
by Servers**

*We do not measure all workloads.  
But, our capacity measure  
accounts for all servers and we  
assume the workloads we do  
measure are representative of  
the whole*

**Measured  
By**

**Assumptions:**

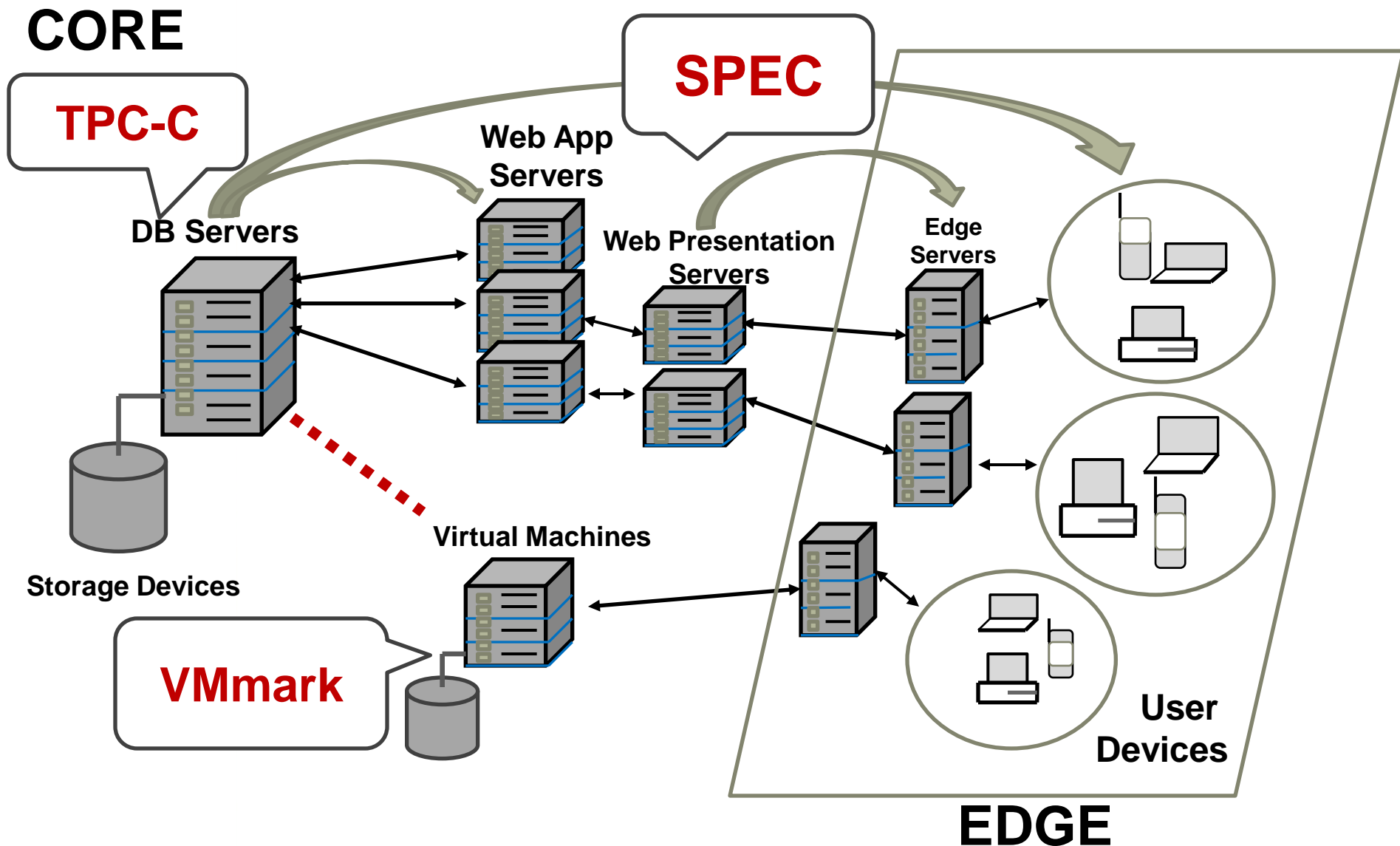
*Simulated workloads represent actual  
workloads*

*Estimating max server processing  
capacity for multiple workloads  
yields meaningful upper limit*

**Industry  
Standard  
Server  
Benchmarks**

**Where Each  
Benchmark  
Simulates One  
or More  
Enterprise  
Workloads**

# Measurement Points

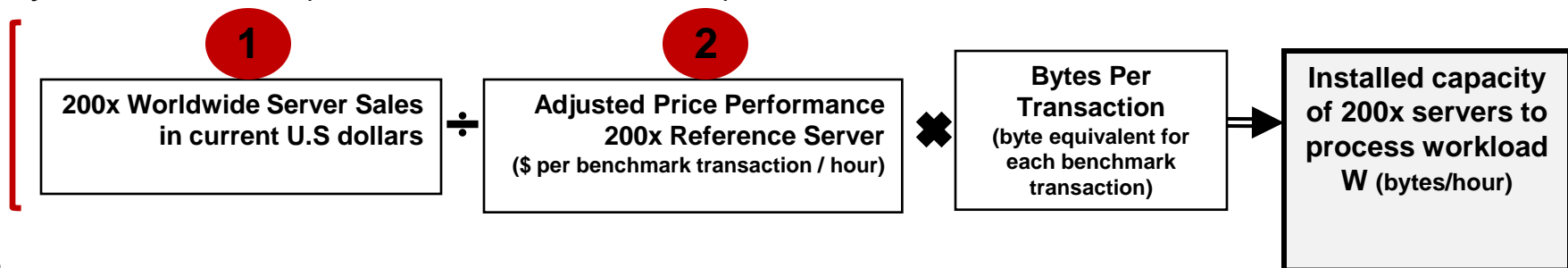




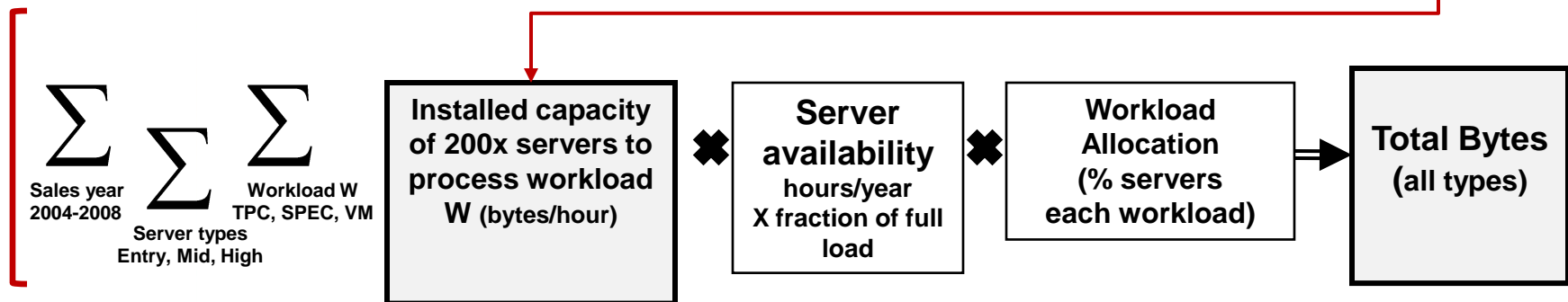


# Modeling Approach

- 1 By sales year 2004-2008  
By server type (entry-level, midrange, high-end)  
By workload **W** (TPC, SPEC, VMmark)



- 2 Summing over sales, server types and workloads





1

# World Server Sales 2004-2008

**Annual World Server Sales  
2004-2008**  
Current U.S. Dollars  
(in billions)

Server Class	2004	2005	2006	2007	2008	Total
Entry-level	\$24.4	\$27.3	\$28.5	\$30.8	\$29.3	\$140.5
Midrange	\$12.8	\$12.8	\$12.2	\$12.6	\$11.7	\$62.3
High-end	\$12.2	\$11.6	\$12.0	\$11.6	\$12.2	\$59.9
<b>Total</b>	<b>\$49.5</b>	<b>\$51.8</b>	<b>\$52.8</b>	<b>\$55.1</b>	<b>\$53.3</b>	<b>\$262.7</b>


Source:

HMI? 2010. Data compiled from IDC Quarterly Server Tracking Reports, 2004-2009.

# COST ANALYSIS

## 2

- We used the detailed TPC-C pricing sheets to break down the costs of the System Under Test (SUT) into the cost of the Server Hardware and other components
- To accord with our server sales numbers, we recalculated price performance using only the costs of the server hardware (current dollars)
- Further cost performance (storage, network) analyses are possible and planned

	PowerEdge 2900			TPC-C 5.9		
				TPC Pricing 1.2		
				Report Date		
				June 16, 2008		
				Revision Date		
				Availability Date		
				June 16, 2008		
Description	Part Number	Price Source	Unit Price	Qty	Extended Price	3 yr. Maint. Price
<b>Server Hardware</b>						
PE2900, QC XEON X5440, 2X6MB, 2.83GZ	223-4506	1	1,253.00	1	\$1,253.00	\$578.00
& 2 Broadcom NICs						
32GB 667MHz(8x4GB), 2R	311-6327	1	3,118.00	1	\$3,118.00	
PERC6/i, Integrated	341-5699	1	\$299.00	1	\$299.00	
PERC6/E SAS RAID, 2X4 EXTERNAL	341-5842	1	\$799.00	3	\$2,397.00	
DELL E157FP, 15 IN, 15.0 VIS	320-5090	1	\$189.00	1	\$189.00	
				<b>Subtotal</b>	<b>\$7,256.00</b>	<b>\$578.00</b>
<b>Server Storage</b>						
PV MD1000, RACK, 3U, 15 BAY, LBZL	222-2299	1	2,480.00	6	\$14,880.00	\$9,888.00
SINGLE ENCL MGT MODULES, SAS/SATA	420-5927	1	\$345.83	6	\$2,074.98	
SAS Cable, 1M, MD1000	310-7082	1	\$30.00	6	\$180.00	
73GB, 3GBPS, SAS, 3.5IN, 15K	341-3023	1	\$299.00	98	\$29,302.00	
42U Rack, CUST	340-4896	1	\$239.99	1	\$239.99	
				<b>Subtotal</b>	<b>\$46,676.97</b>	<b>\$9,888.00</b>
<b>Server Software</b>						
Oracle Database 11g Standard Edition One, Per Processor		2	\$2,498.00	1	\$2,498.00	
Unlimited Users, 3 years						
Windows Server 2003 Standard x64 Server	420-7118	1	\$799.00	1	\$799.00	
Microsoft Problem Resolution Services		3	\$245.00	1		\$245.00
Oracle Premium Support, 3 years		2	\$1,099.00	3		\$3,297.00
				<b>Subtotal</b>	<b>\$3,297.00</b>	<b>\$3,542.00</b>

**Server Hardware:**  
**\$7,256**

**Storage System:**  
**\$46,677**

**SUT Test System:**  
**\$65,910**

1	\$760.00	\$472.00
1	\$599.00	
1	\$448.00	
1	\$0.00	
1	\$59.00	
1	\$99.00	
1	\$149.00	
<b>Subtotal</b>	<b>\$2,114.00</b>	<b>\$472.00</b>
1	\$799.00	
1	\$250.00	
<b>Subtotal</b>	<b>\$1,049.00</b>	<b>\$0.00</b>
3	\$3.93	
<b>Subtotal</b>	<b>\$3.93</b>	<b>\$0.00</b>
<b>16% discount</b>	<b>(\$8,967.52)</b>	
<b>Total</b>	<b>\$51,429.38</b>	<b>\$14,480.00</b>

<p>*All hardware items from Dell(1) are discounted 16% based on total dollar volume of this configuration.</p> <p>Price Source: 1=Dell, 2=Oracle, 3=Microsoft, 4=Kalron</p> <p>Pricing may be verified by calling 1-800-BUY-DELL and referencing quote # 432979497 as a complex quote.</p> <p><b>Audited by Lorna Livingtree, Performance Metrics Inc.</b></p> <p>Prices used in TPC benchmarks reflect the actual prices a customer would pay for a one-time purchase of the stated components. Individually negotiated discounts are not permitted. Special prices based on assumptions about past or future purchases are not permitted. All discounts reflect standard pricing policies for the listed components. For complete details, see the pricing sections of the TPC benchmark specifications. If you find that the stated prices are not available according to these items, please inform the TPC at <a href="mailto:pricing@tpc.org">pricing@tpc.org</a>.</p>	<b>Three-Year Cost of Ownership:</b>		<b>\$65,910</b>	<b>USD</b>
	<b>TPC-C Throughput:</b>		<b>97,083.53</b>	<b>tpmC</b>
	<b>Price/Performance:</b>		<b>\$0.68</b>	<b>tpmC/USD</b>





# World Server Capacities

ENTRY-LEVEL SERVERS	Reference Year	tpmC	Price/tpmC	Total System Cost	Server Hardware (SH) Cost	Adjusted SH_Price/tpmC	Core Capacity tpmC
Dell PowerEdge 2900	2008	97,083	\$0.68	\$65,910	\$7,256	\$0.07	3.92E+11
Dell PowerEdge 2900	2007	69,564	\$0.91	\$63,080	\$11,658	\$0.17	1.84E+11
Dell PowerEdge 2900	2006	65,833	\$0.98	\$64,512	\$9,839	\$0.15	1.91E+11
Dell PowerEdge 2800	2005	38,622	\$0.99	\$38,028	\$10,771	\$0.28	9.81E+10
Dell PowerEdge 2850	2004	26,410	\$1.53	\$40,170	\$7,993	\$0.30	8.08E+10
Estimated	2003					\$0.33	7.34E+10

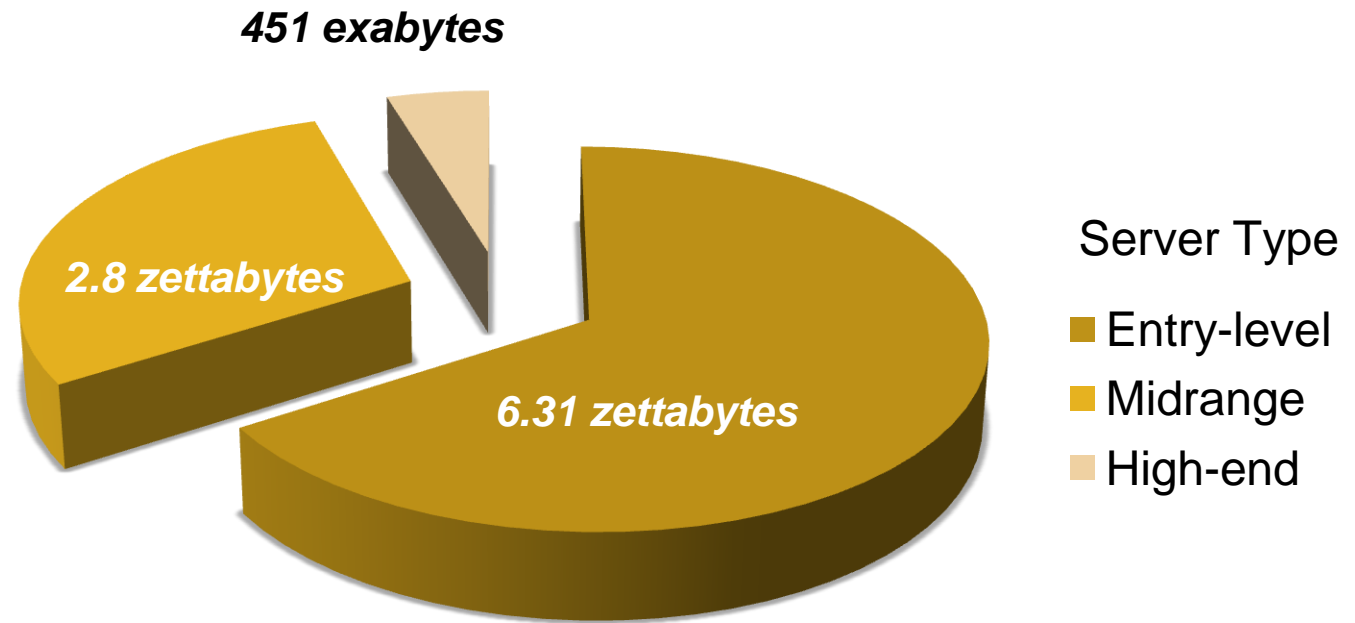
MIDRANGE SERVERS	Reference Year	tpmC	Price/tpmC	Total System Cost	Server Hardware (SH) Cost	Adjusted SH_Price/tpmC	Core Capacity tpmC
HP ProLiant DL585 G5	2008	402,234	\$1.26	\$502,836	\$43,634	\$0.11	1.08E+11
HP ProLiant DL580 G5	2007	407,079	\$1.71	\$694,335	\$70,560	\$0.17	7.33E+10
HP ProLiant ML370 G5	2006	240,737	\$1.85	\$443,443	\$83,504	\$0.35	3.53E+10
HP ProLiant DL585	2005	130,623	\$2.80	\$364,539	\$87,372	\$0.67	1.92E+10
HP ProLiant DL580 G2	2004	95,163	\$2.93	\$278,114	\$72,151	\$0.76	1.69E+10
Estimated	2003					\$0.87	1.47E+10

HIGH-END SERVERS	Reference Year	tpmC	Price/tpmC	Total Price	Server Hardware (SH) Cost	Adjusted SH_Price/tpmC	Core Capacity tpmC
IBM System x3950 M2	2008	841,809	\$3.46	\$2,911,484	\$1,020,576	\$1.21	1.01E+10
IBM System p 570	2007	404,462	\$3.50	\$1,417,121	\$625,499	\$1.55	7.50E+09
IBM System p5 570	2006	203,440	\$3.93	\$799,990	\$514,449	\$2.53	4.76E+09
IBM eServer p5 570	2005	194,391	\$5.62	\$1,092,119	\$527,839	\$2.72	4.29E+09
IBM eServer p5 570	2004	371,044	\$5.26	\$1,951,215	\$1,035,538	\$2.79	4.39E+09
Estimated	2003					\$3.35	3.66E+09

# New Way To Measure Capacity

- Aggregate capacity over very different kinds of servers
- Assume companies spend dollars for server capacity efficiently
- Measure capital cost per unit of benchmarked performance
- Then we translate different benchmarks into a common unit: bytes

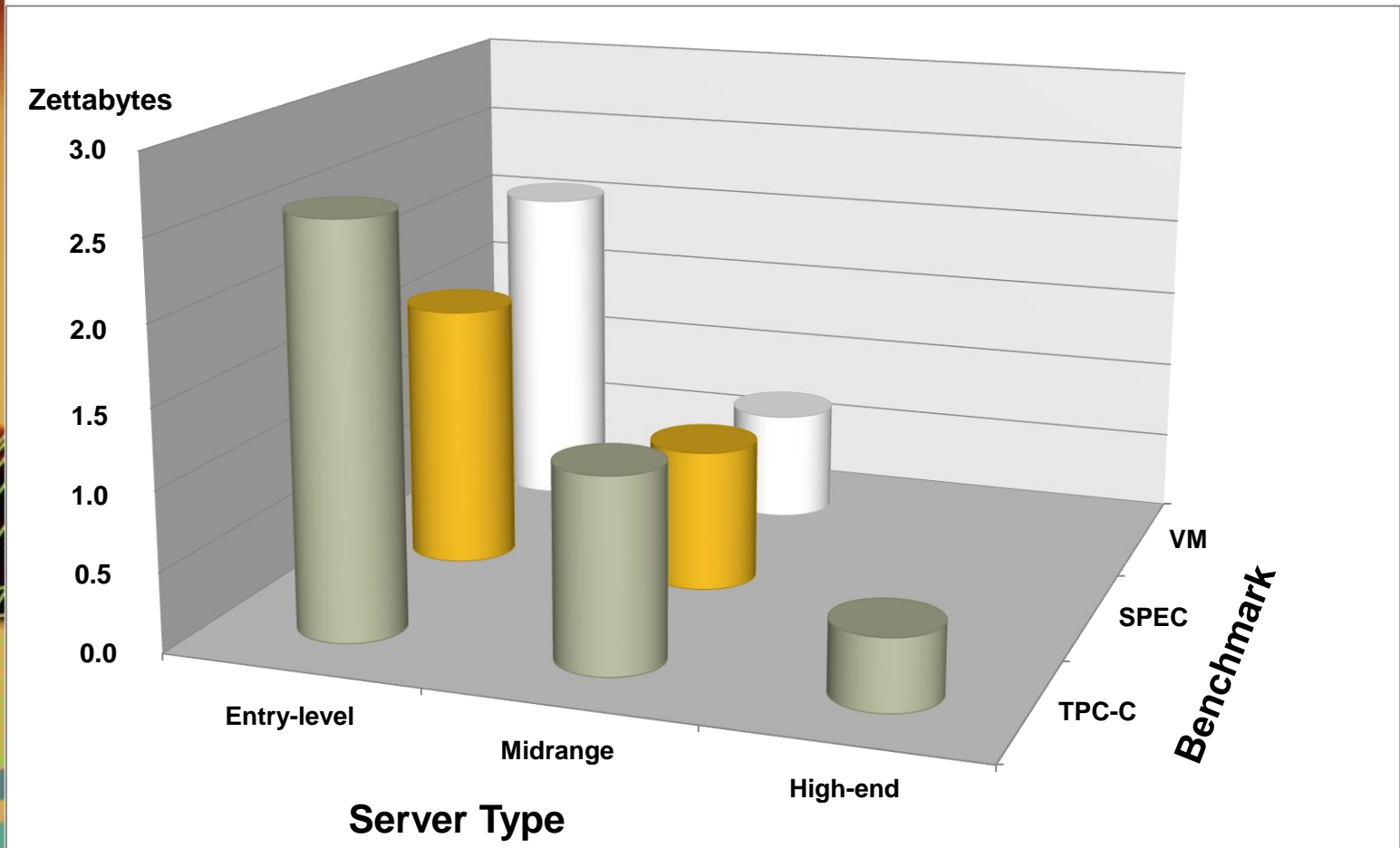
# World Server Summary Information



**IDC/EMC 2010 Total Digital Universe:** 1.2 zettabytes in 2010  
35 zettabytes by 2020

**Lyman/Varian 2003:** 18 exabytes of new information produced annually

# Class Server Contribution to Total



Total = 9.57 zettabytes =  $9.57 \times 10^{21}$  bytes



# Discussion and Implications

- Results expressed in capacity per dollar (not capacity per server)
- Data Intensive Computing Platforms
  - Large memory systems
  - Shared nothing platforms
  - Database machines
- Ongoing and Future Research
  - Center for Large Scale Data Systems Research (CLDS)

# Center for Large-Scale Data Systems (CLDS): What?

- A center dedicated to the study of technical, management, and economic issues related to large-scale data systems
  - Investigate / study architectures and software systems for cloud storage systems
  - Analyze / develop relevant benchmarks and cost analysis
    - Build upon early work by the HMI? project
  - Create forums for exchange of ideas
    - Among different industry segments, e.g. applications versus infrastructure
    - Among entities dealing with different types of data, e.g. enterprise warehouses, data streams, scientific data, healthcare

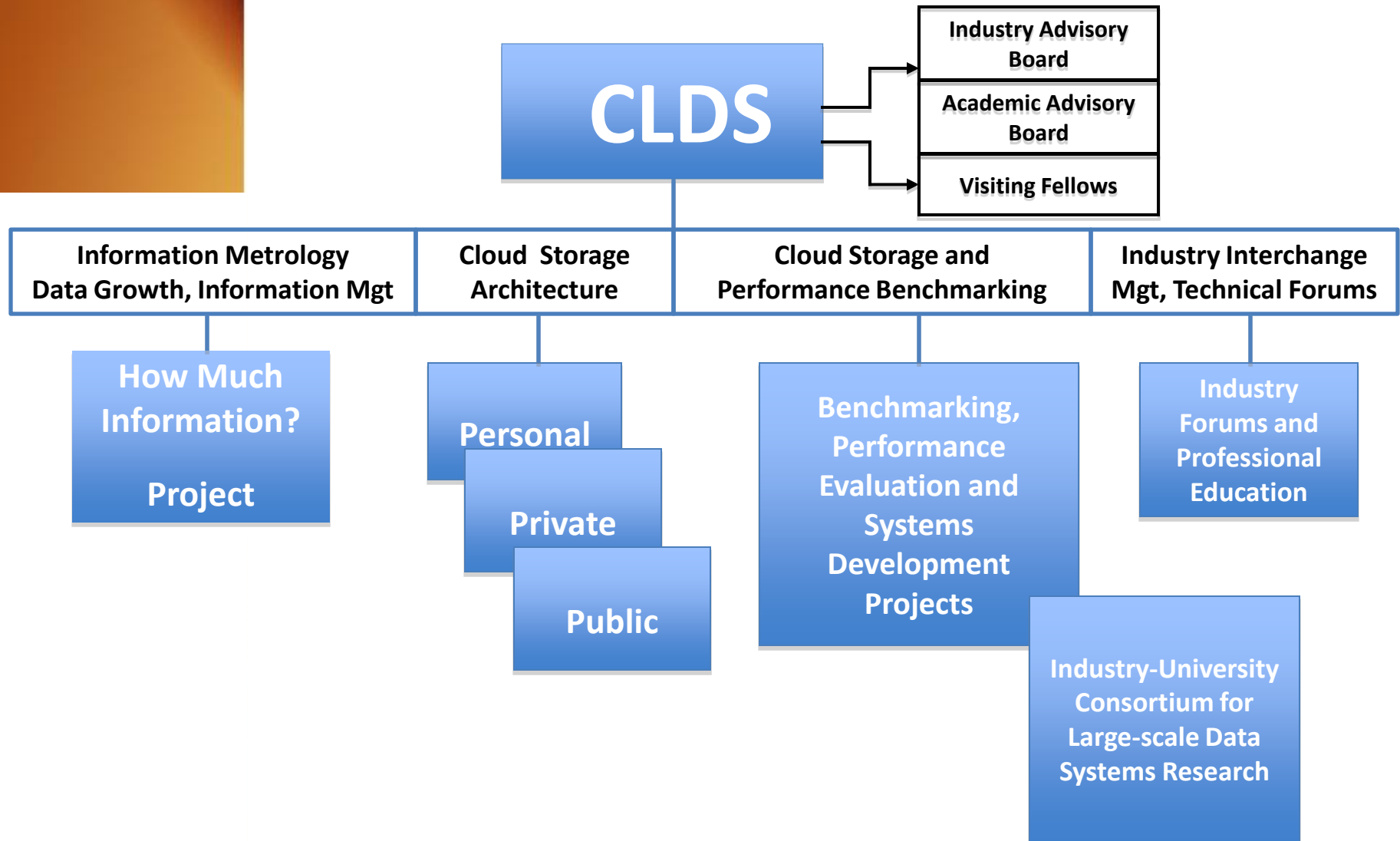
# CLDS Rationale: Why?

- Dealing with the *data deluge*
- Addressing the *fourth paradigm*: data-intensive science and decision making
- Understanding the paradigm shift occurring with *cloud computing*, specifically for large-scale data
- Need for a research and industry forum that can step back, assess situation, visualize current and future trends, provide insights and offer solutions

# CLDS Structure: How?

- Participate in projects on
  - Information Metrology, Data Growth, and Information Management
    - Build upon on-going How Much Information? project
  - Cloud Storage Architecture
  - Cloud Storage Performance Benchmarking
    - Build upon on-going cloud-related research at SDSC
- Obtain insights into application requirements
  - For different application domains
- Participate in industry forums and professional education
  - Serve on Industry Advisory Board, Fellows Program
  - Help design and develop professional ed programs

# CLDS Center Structure





# Example CLDS Projects

- HMI?: How Much Information?
  - HMI Consumer, HMI Enterprise reports
  - HMI enterprise server study – adding storage and network
  - HMI company projects
    - Data Mobility and Information Value
    - CIO Interview Project: Managing the Challenges and Opportunities in “Big Data”
    - Pilot Projects on Managing and Exploiting Information Value
- Cloud Storage Architecture, Benchmarking & Performance
  - Management and technical systems research on personal, private, and public cloud storage architectures
  - Cloud storage systems evaluation and performance benchmarking, including development and testing of new benchmarks
  - Cloud storage capacity and performance evaluation by industry vertical and by enterprise workload

# For More Information

- Chaitanya Baru [baru@sdsc.edu]
- Jim Short [jshort@ucsd.edu]

Director, Industry Relations, SDSC

- Ron Hawkins [rhawkins@sdsc.edu]