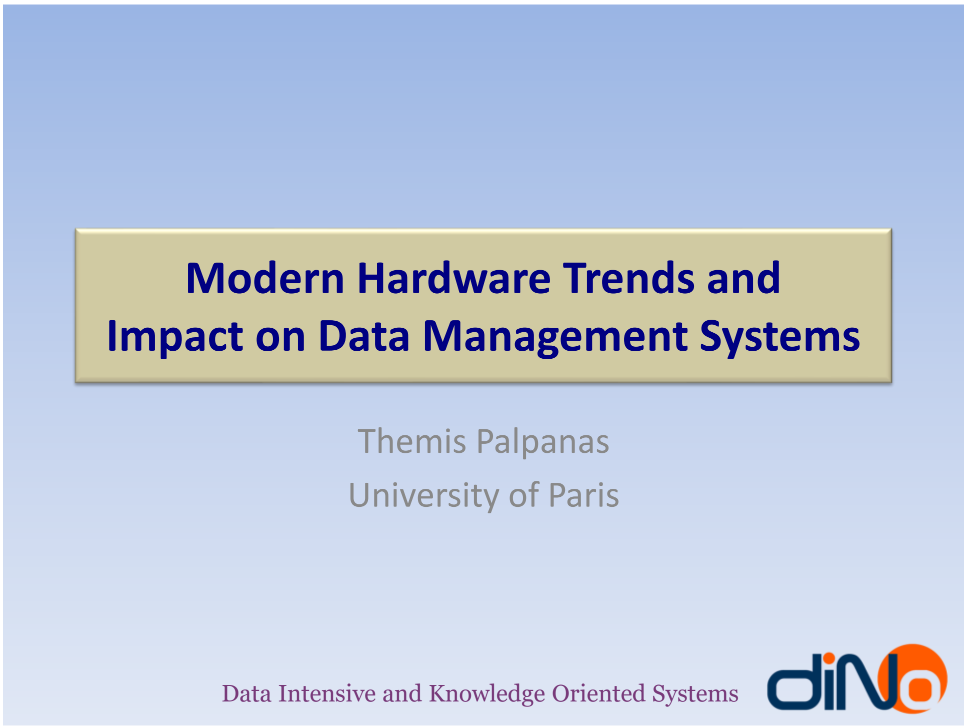


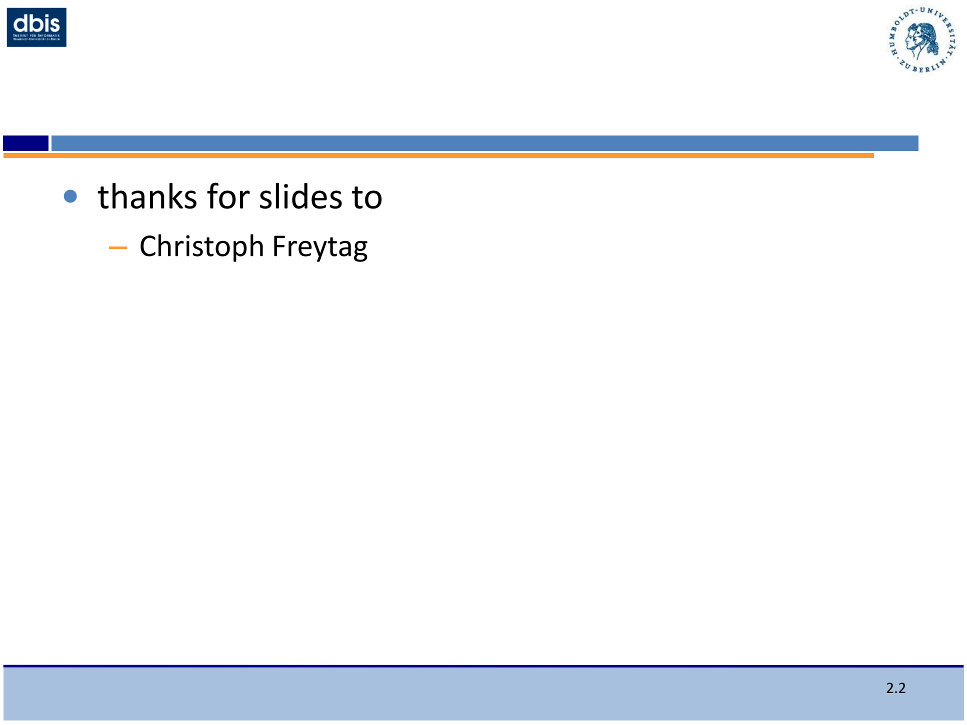
Modern Hardware Trends and Impact on Data Management Systems

Themis Palpanas

University of Paris

Data Intensive and Knowledge Oriented Systems

1



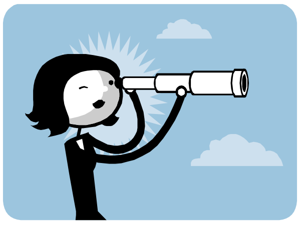
- thanks for slides to
  - Christoph Freytag

2.2

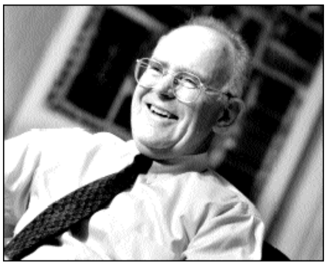
2

Overview

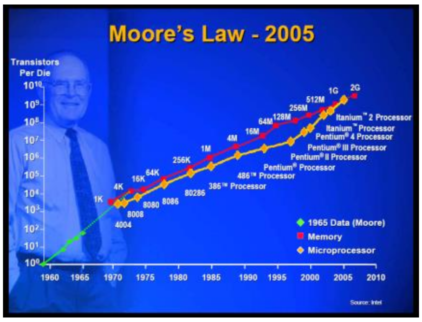
- Hardware aspects
  - Moore‘s Law
  - New developments in HW
- What’s does it mean for
  - Algorithms/Data Structures
  - DBMS architecture
- Questions & Challenges
- Future developments



Technology Trends: Microprocessor Capacity



Gordon Moore (co-founder of Intel) predicted in 1965 that the transistor density of semiconductor chips would double roughly every 18 months.



Source: Intel

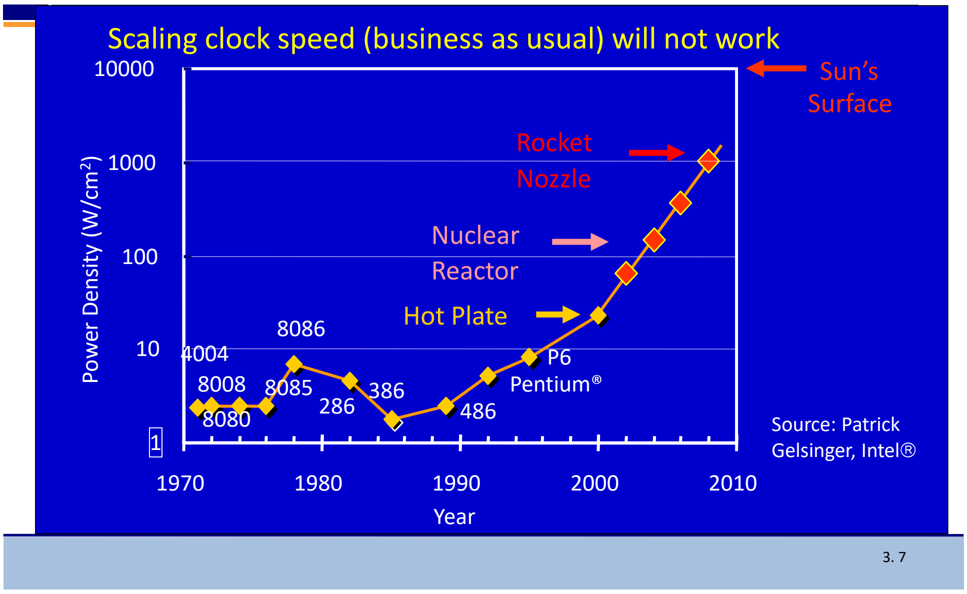
2X transistors/Chip Every 1.5 – 1.8 years
Called “Moore’s Law”

Microprocessors have become smaller, denser, and more powerful.

Moore‘s Law - the problem (1)

- # of transistors on-chip doubles every 18 months
  - So much of innovation was possible only because we had transistors
  - Phenomenal 58% performance growth every year
- Moore’s Law faced a danger around 2000
  - Power consumption is too high when clocked at multi-GHz frequency
  - it is proportional to the number of switching transistors
- Wire delay doesn’t decrease with transistor size

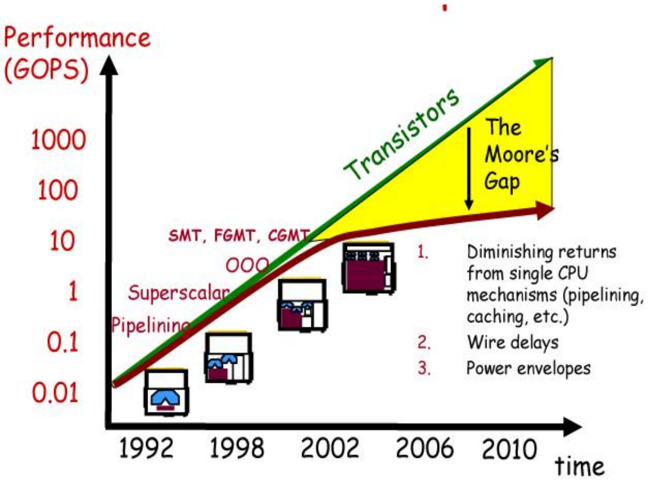
Changes in power density



Scaling clock speed (business as usual) will not work

Source: Patrick Gelsinger, Intel®

Moore‘s Law – the problem (2)



Moore‘s Law - the problem (3)

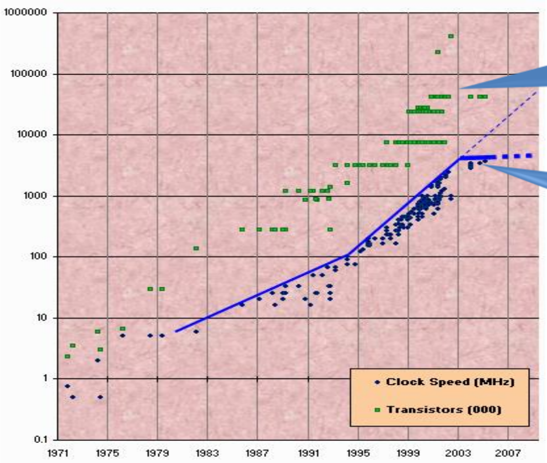
- Hardware for extracting ILP (Instruction Level Parallelism) has reached the point of diminishing return
  - Need a large number of in-flight instructions
  - Supporting such a large population inside the chip requires power-hungry delay-sensitive logic and storage
- Verification complexity is getting out of control
- How to exploit so many transistors?
  - Must be a de-centralized design which avoids long wires

Moore‘s Law – the problem (4)

Pentium 3	Pentium 4
1 GHz	1.4 GHz
Year 2000	Year 2000
0.18 micron	0.18 micron
28M transistors	42M transistors
343 (Specint 2000)	393 (Specint 2000)

Transistor count increased by 50%
Performance increased by only 15%

New developments in HW



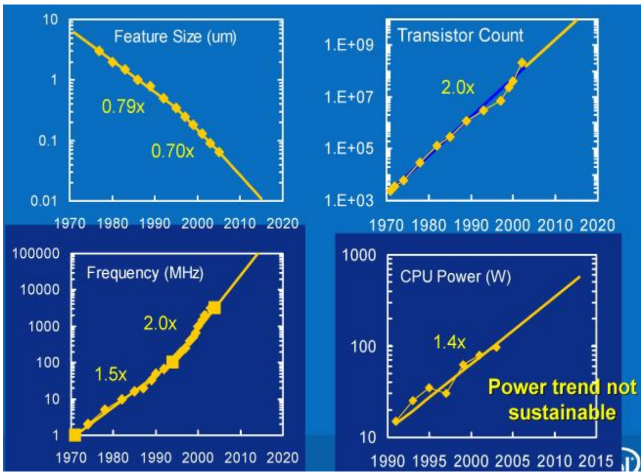
Transistor density still rising

Clock speed isn’t

Transistors are used for parallelism:
Multicore processors

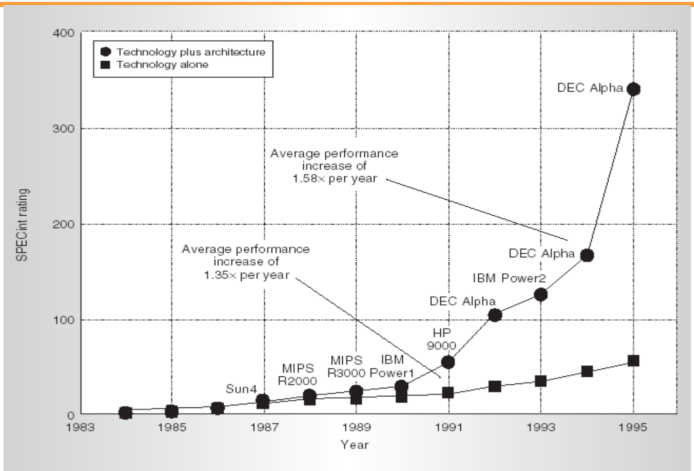
Source: Sutter, The Free Lunch is over

New developments in HW



From http://www.cs.jhu.edu/~spaa/2006/SPAA06-Lowney.pdf

CPU development

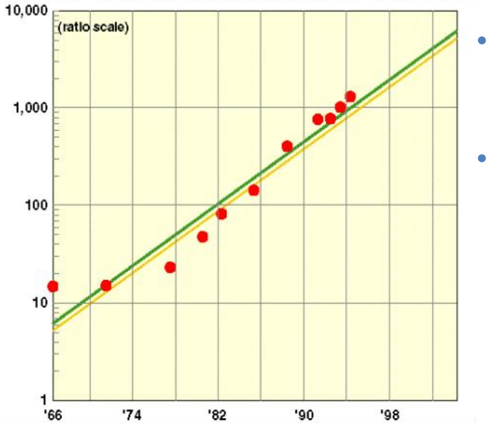


Slightly outdated data

Problem in chip production

Manufacturing costs and yield problems limit use of density

Cost of semiconductor factories in millions of 1995 dollars

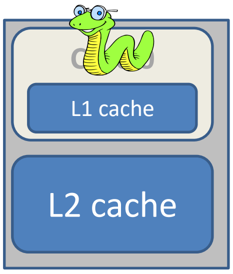


- Moore’s (Rock’s) 2nd law:
  - Fabrication costs go up
  - Yield (% usable chips) drops
- Parallelism can help
  - Smaller, simpler processors are easier to design and validate
  - Can use partially working chips:
  - E.g., Cell processor (PS3) is sold with 7 out of 8 “on” to improve yield

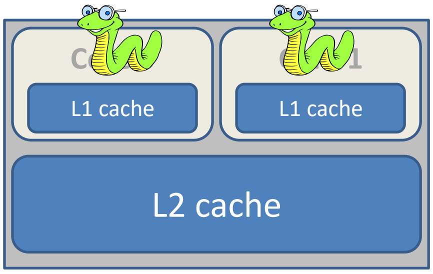
Physical measures and constraints

- Reducing power with voltage scaling
  - Power = Capacitance x Voltage^2 x Frequency
  - Frequency ~ Voltage in “region of interest”
  - Power ~ Voltage^3
- Example: 10% of reduction in voltage yields
  - 10% reduction in frequency
  - 30% reduction in power
  - Less than 10% reduction in performance

Conventional vs. Multicore



Conventional processor
- Single core
- Dedicated caches
- One thread at a time

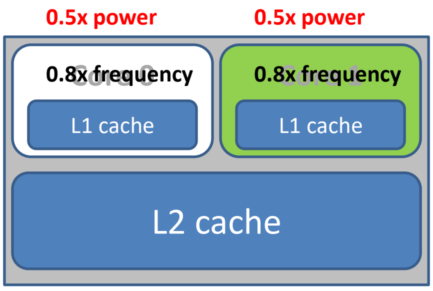


Multicore processors
- At least ***two cores***
- Shared caches
- Many threads simultaneously

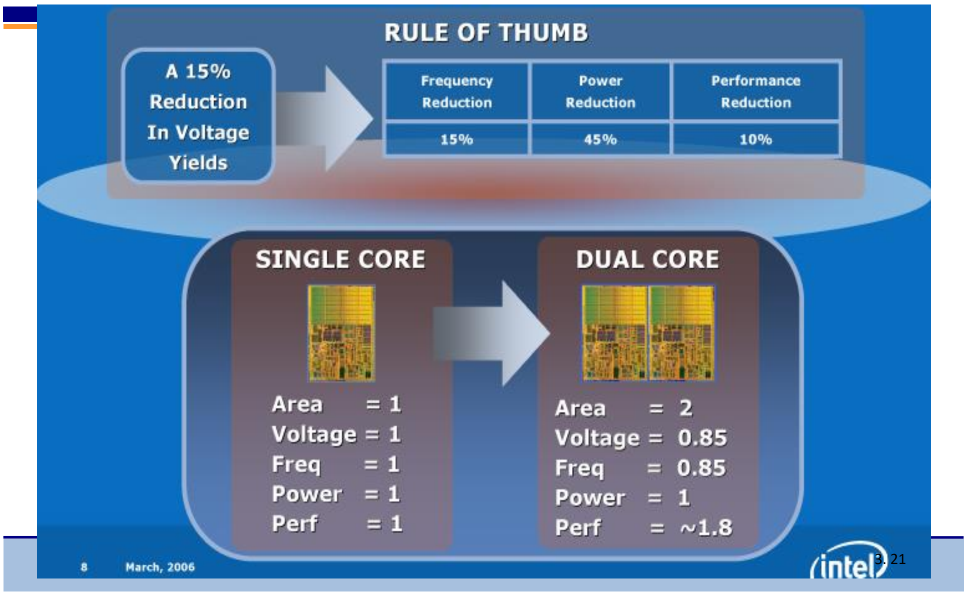
Superior Performance/Watt

- Example:
  - Reduce CPU clock frequency by 20%
  - Power consumption reduces by 50%!

- Put two 0.8 frequency cores on the same chip
- Get 1.6 times the computation at the same power consumption



Reduce Voltage – double core



RULE OF THUMB

A 15% Reduction In Voltage Yields

Frequency Reduction	Power Reduction	Performance Reduction
15%	45%	10%

SINGLE CORE

Area = 1
Voltage = 1
Freq = 1
Power = 1
Perf = 1

DUAL CORE

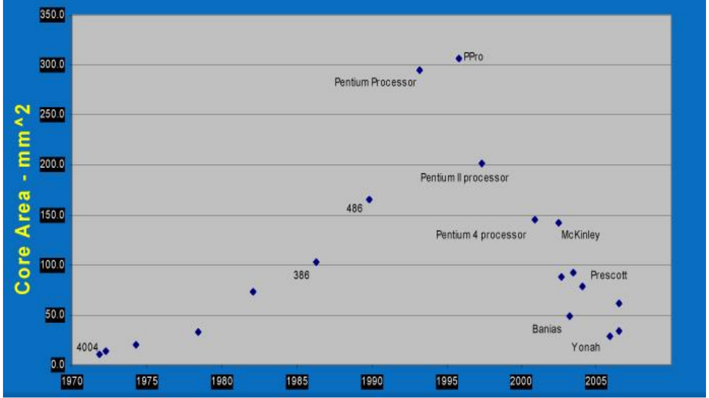
Area = 2
Voltage = 0.85
Freq = 0.85
Power = 1
Perf = ~1.8

8 March, 2006

Area development

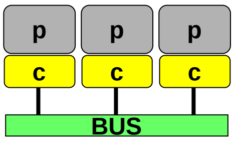
Core Area (with L1 caches) Trend

Die area/core shrinking after peaking with PPro

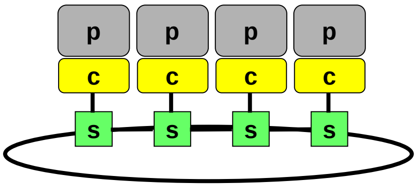


Interconnect Options

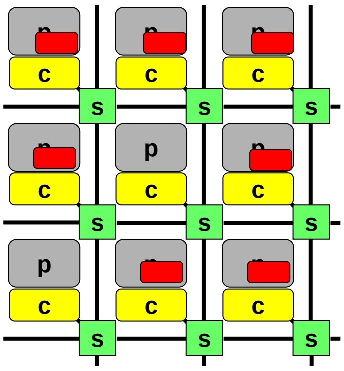
Bus Multicore



Ring Multicore

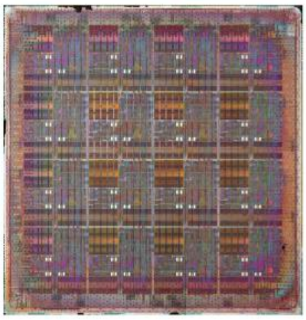


Mesh Multicore



Packet routing through switches

Multicore CPUs: MIT Raw Project

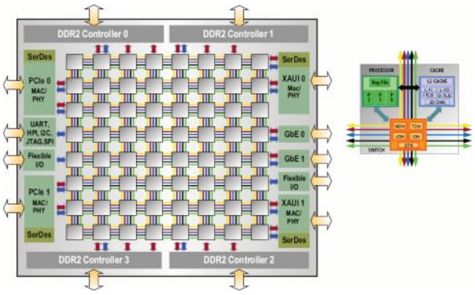


- 16 cores
- Year 2002
- 0.18 micron
- 425 MHz
- IBM SA27E std. cell
- 6.8 GOPS

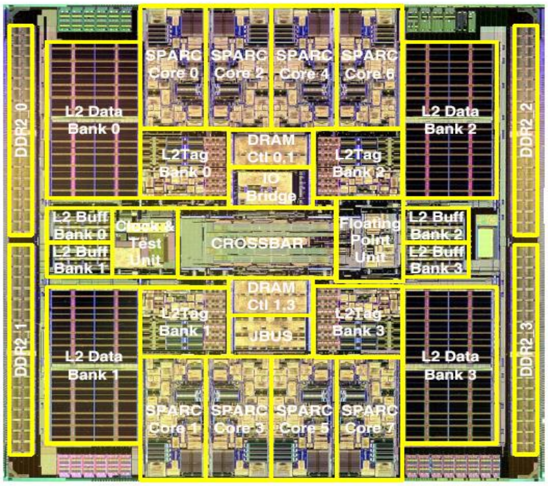
Please see for more information:
http://groups.csail.mit.edu/cag/raw/

Tilera – 64 Core CPU

- Tiling architecture
  - Regular tiling structure
  - Mesh interconnect
- Start-Up from MIT
  - Anant Agarwal
  - From Raw Project
- News October 10,2012:
  - Tile-Gx9 chip can tackle at least nine tasks at the same time, and it does so while consuming less than 10 watts of power



Multicore CPU: Niagara (SUN/Oracle)



Features:

- Eight 64b Multithreaded SPARC Cores
- Shared 3MB L2 Cache
- 16KB ICache per Core
- 8KB DCache per Core
- Four 144b DDR-2 DRAM Interfaces (400 MTs)
- 3.2GB/s JBUS I/O
- Crypto: Public Key (RSA)
- Extensive RAS

Technology:

- 90nm CMOS Process
- 9LM Copper Interconnect
- Power: 63 Watts @ 1.2GHz
- Die Size: 378mm^2
- 279M Transistors
- Package: Flip-chip ceramic LGA (1933 pins)

Development of Niagara/SPARC T-Series

- Developed by Sun/Oracle

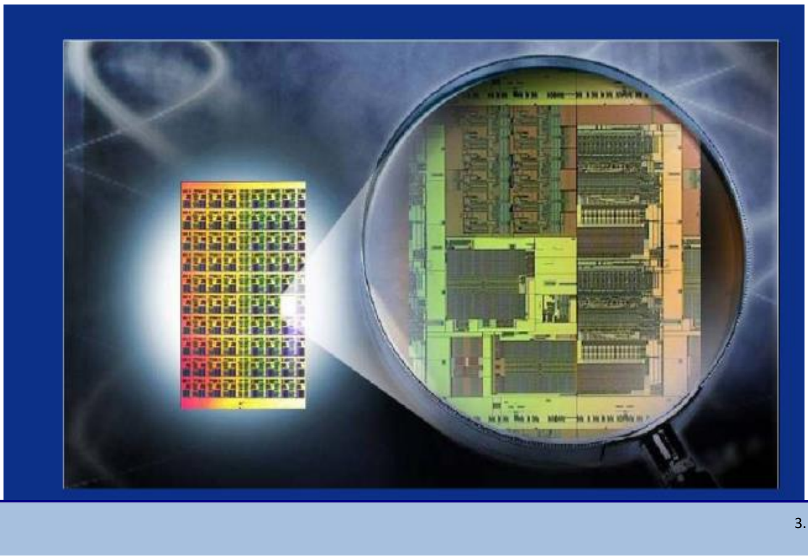
	Year - Release	# of Cores	Clock Rate (GHz)	Threads per Core	Size L1/L2 Cache	Size L3 Cache
UltraSPARC T1	11/2005	4/6/8	1.0 – 1.4	4	L1: 16Kb (I)/8kB (D) pC L2: 3 MB 8shared)	
UltraSPARC T2	10/2007	4/6/8	1.2 – 1.6	Up to 8	L1 16kB (I)/8kB (D) L2: 4MB (shared)	
UltraSPARC T3	10/2010	8/16	1.65	8	L1: 16KB(I)+8KB(D) pC L2: 6MB (shared)	
UltraSPARC T4	Q4/2011	8	2.8 – 3	8	L1: 16kB(I)/16KB (D) pC L2: 128kB pC	8 MB (shared)
UltraSPARC T5	2013	16		8	L1: 16kB (I)/16KB (D) pC L2: 128kB pC	8Mb (shared)

UltraSPARC T3: http://www.spec.org/jEnterprise2010/results/res2010q3/jEnterprise2010-20100825-00014.txt

UltraSPARC T4: http://en.wikipedia.org/wiki/SPARC_T4

UltraSPARC T5: http://www.theregister.co.uk/2012/09/04/oracle_sparc_t5_processor/

Intel Polaris (80 Core CPU)

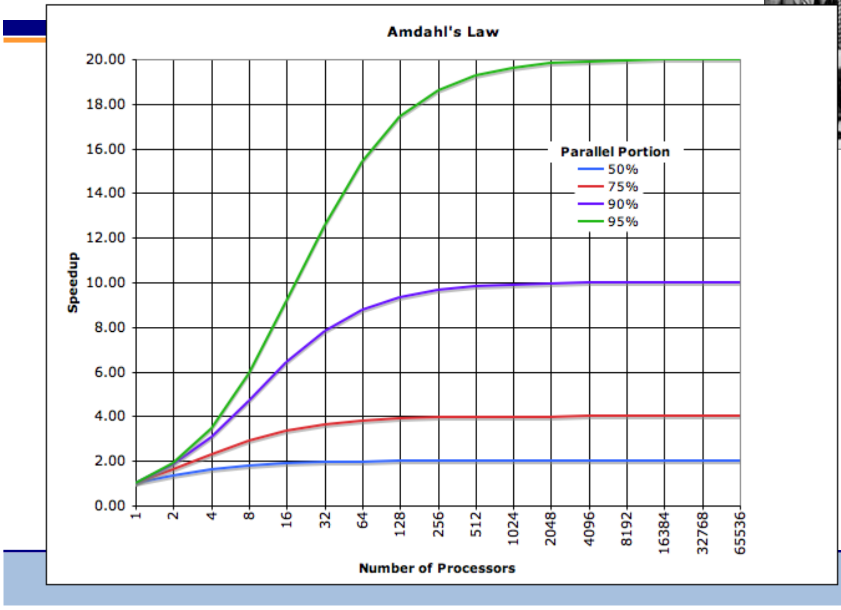


Amdahl‘s Law

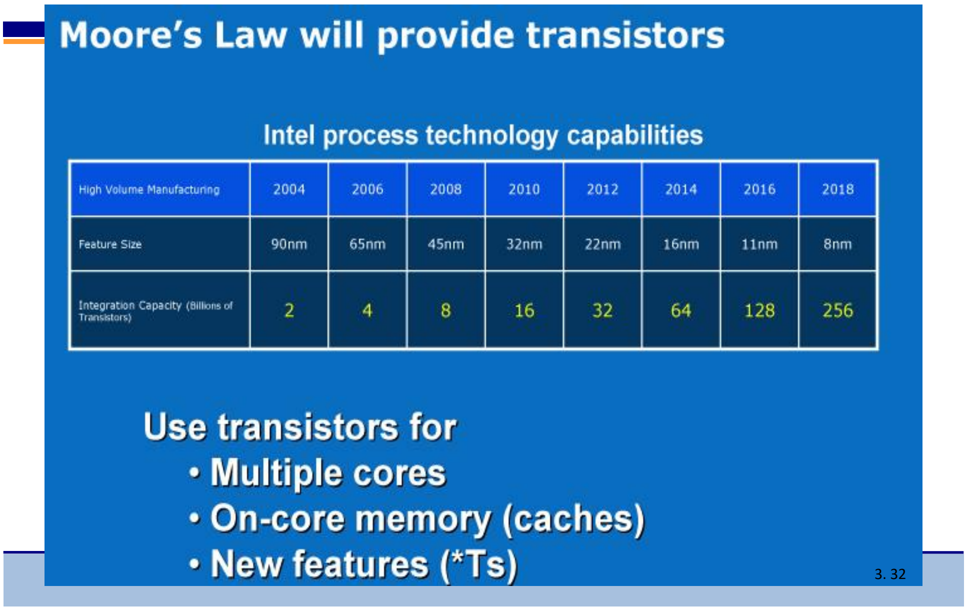
- Assumptions:
  - Let p be the part of a program that is parallelizable
  - (1-p) is the part that can only be executed sequentially
  - Let N be the number of available cores/CPUs
- Then the speedup S can be computed as

$$S = \frac{1}{(1-p) + \frac{p}{N}}$$

Amdahl‘s Law



Future of Multicore CPUs



Moore’s Law will provide transistors

Intel process technology capabilities

High Volume Manufacturing	2004	2006	2008	2010	2012	2014	2016	2018
Feature Size	90nm	65nm	45nm	32nm	22nm	16nm	11nm	8nm
Integration Capacity (Billions of Transistors)	2	4	8	16	32	64	128	256

Use transistors for

- Multiple cores
- On-core memory (caches)
- New features (*Ts)

Outlook

- With a doubling of cores every 18 months, 100s to 1000s of powerful threads on a chip soon

Year	2008	2011	2014	2017
# Cores	4	16	64	256
# Threads	16	64	256	1024

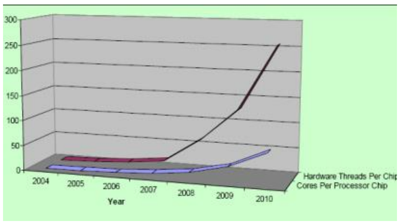
Memory System Performance

Memory Access Latency in nanoseconds

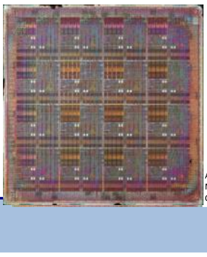
	L1	L2	Main Memory	Random Memory
Intel	1.1290	5.2930	118.7	150.3
AMD	1.0720	4.3050	71.4	173.8

Multicore CPU is here today

- ☐ Multi core CPUs
  - Most people know …
  - Little understanding how to use…
- ☐ Facts
  - Up to 64-128 cores per CPU
  - More than 32 MB of (shared?) L2-cache
- ☐ Must think differently for SW



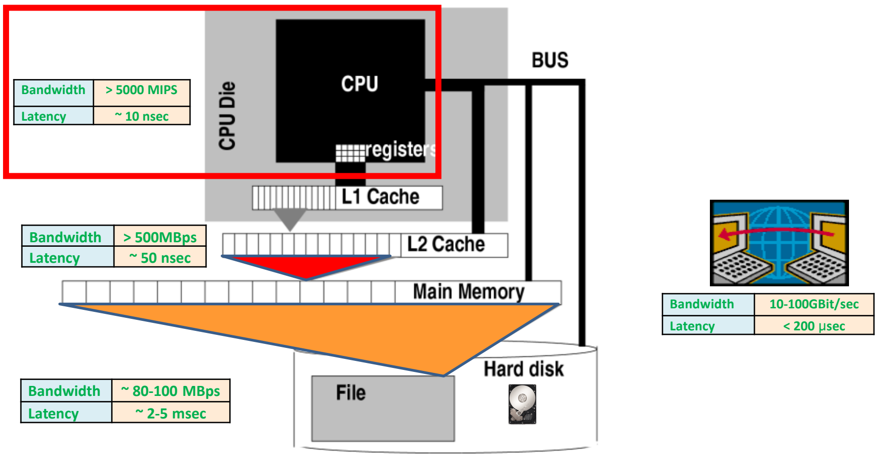
Source: The Impact of Multicore on Math Software … : Workshop on Edge Computing Using New Commodity Architectures (EDGE), NC, Chapel Hill 2006



- 16 cores
- Year 2002
- 0.18 micron
- 425 MHz
- IBM SA27E std. cell
- 6.8 GOPS

Agarwal, A. (2006). The Why, How and When of Multicore. *EDGE Workshop*. University of North Carolina at Chapel Hill, 2006

Trends we currently see…



Bandwidth	> 5000 MIPS
Latency	~ 10 nsec

Bandwidth	> 500MBps
Latency	~ 50 nsec

Bandwidth	~ 80-100 MBps
Latency	~ 2-5 msec

Bandwidth	10-100GBit/sec
Latency	< 200 µsec

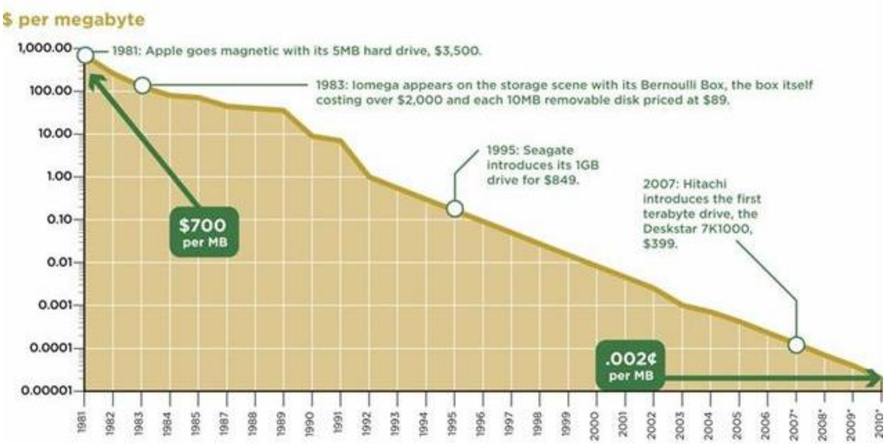
Changing technology: Flash disk

- Characteristics
  - 2012: ~2TB - Cost about $400
  - Less power consumption !!

Device	Sequential	Random 8KB	Price $	Power	iops/$	iops/watt
SCSI 15k rpm	75 MBps	200 iops	500$	15 watt	0.5	13
SATA 10k rpm	60 MBps	100 iops	150$	8 watt	0.7	12
Flash- read	53 MBps	2,800 iops	400$	0.9 watt	7.0	3,100
Flash - write	36 MBps	27 iops	400$	0.9 watt	0.07	30

Gray, J., & Fitzgerald, B. (2007), FLASH Disk Opportunity for Server-Applications, from http://research.microsoft.com/~Gray/papers/FlashDiskPublic.doc; Jan 2007; Retrieved March 8, 2007

Changing cost of storage

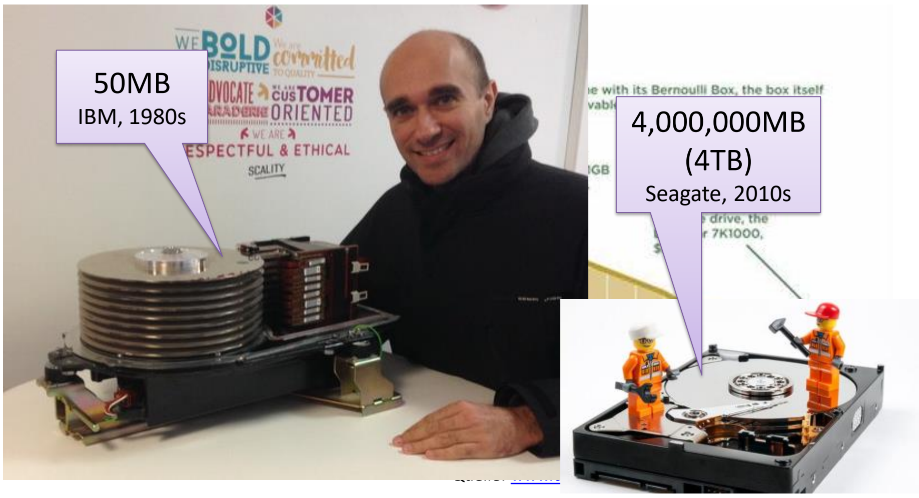


Quelle: www.deepspar.com/images/Storage_Cost.jpg

Changing cost of storage

pspar.com/images/Storage_Cost.jpg

Changing cost of storage



50MB
IBM, 1980s

4,000,000MB
(4TB)
Seagate, 2010s

WE BOLD DISRUPTIVE
We are committed TO QUALITY
ADVOCATE
WE ARE CUSTOMER ORIENTED
WE ARE RESPECTFUL & ETHICAL
SCALITY
e with its Bernoulli Box, the box itself
drive, the
7K1000,

Storage technology in 2011

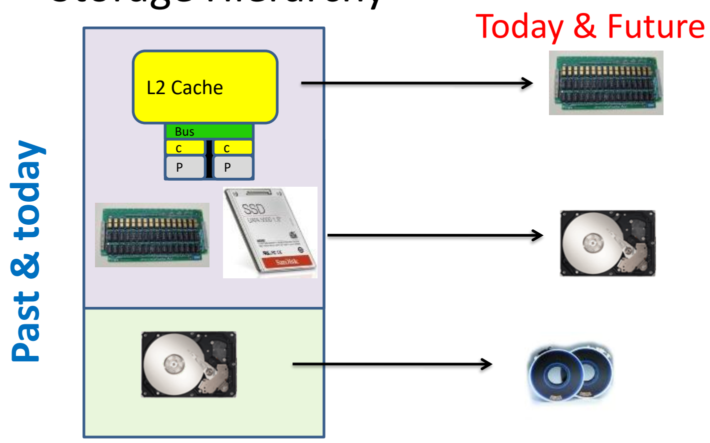
Technology type	Revenues [billion $]	#units shipped [million]	Sold storage size [ExaBytes]
	Samsung, Hynix, Micron 91% market share		
DRAM Memory	31	800	2
	Samsung, Toshiba, Micron, Hynix 99% market share		
NAND Memory	30	4000	20
	> 50 companies		
Solid State Disks	5	17	3
	Western Digital 37%, Seagate 47% , Toshiba 16%		
Hard-Disk-Drive	28	630	350
	LTO-Consortium, IBM, Oracle		
Magnetic Tape	1	27	20

World market – different technologies

Source: https://espace.cern.ch/WLCG-document-repository/Technical_Documents/Technology_Market_Cost_Trends_2012_v23.pdf

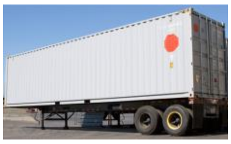
What does this mean for DBMS? (1)

- Storage Hierarchy

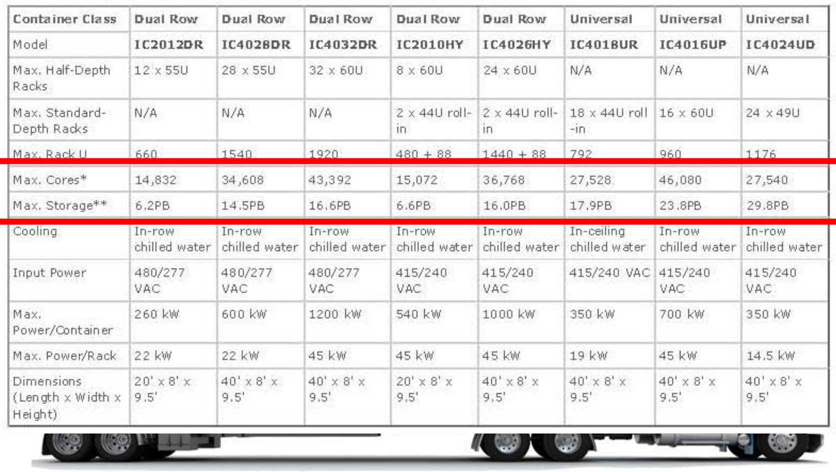


Changing technology: CPU farms

- Example SGI (Silicon Graphics)
  - Before: Rackable Inc.
  - See http://www.sgi.com
- Properties
  - 1200 CPUs
  - 22000 cores
  - 5.4 TB Main memory
  - 7.0 PBytes Disk storage
  - Only Need power & Internet access & water

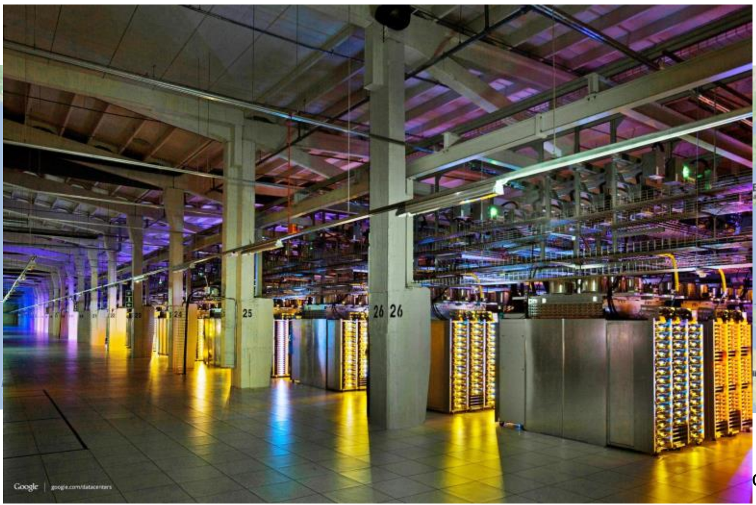


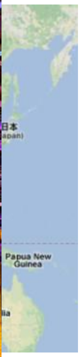
Trends



Container Class	Dual Row	Dual Row	Dual Row	Dual Row	Dual Row	Universal	Universal	Universal
Model	IC2012DR	IC4028DR	IC4032DR	IC2010HY	IC4026HY	IC4018UR	IC4016UP	IC4024UD
Max. Half-Depth Racks	12 x 55U	28 x 55U	32 x 60U	8 x 60U	24 x 60U	N/A	N/A	N/A
Max. Standard-Depth Racks	N/A	N/A	N/A	2 x 44U roll-in	2 x 44U roll-in	18 x 44U roll -in	16 x 60U	24 x 49U
Max. Rack U	660	1540	1920	480 + 88	1440 + 88	792	960	1176
Max. Cores*	14,832	34,608	43,392	15,072	36,768	27,528	46,080	27,540
Max. Storage**	6.2PB	14.5PB	16.6PB	6.6PB	16.0PB	17.9PB	23.8PB	29.8PB
Cooling	In-row chilled water	In-row chilled water	In-row chilled water	In-row chilled water	In-row chilled water	In-ceiling chilled water	In-row chilled water	In-row chilled water
Input Power	480/277 VAC	480/277 VAC	480/277 VAC	415/240 VAC	415/240 VAC	415/240 VAC	415/240 VAC	415/240 VAC
Max. Power/Container	260 kW	600 kW	1200 kW	540 kW	1000 kW	350 kW	700 kW	350 kW
Max. Power/Rack	22 kW	22 kW	45 kW	45 kW	45 kW	19 kW	45 kW	14.5 kW
Dimensions (Length x Width x Height)	20' x 8' x 9.5'	40' x 8' x 9.5'	40' x 8' x 9.5'	20' x 8' x 9.5'	40' x 8' x 9.5'	40' x 8' x 9.5'	40' x 8' x 9.5'	40' x 8' x 9.5'

Google‘s Data Centers





cations/

Example: Google – server farms

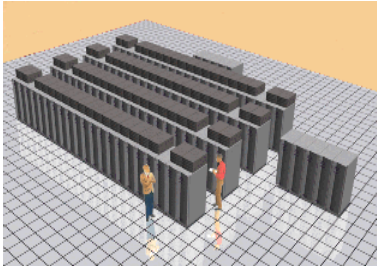
- Movie:
  - http://www.cbsnews.com/video/watch/?id=50133304n

  (from http://tech.slashdot.org/comments.pl?sid=3191691&cid=41680953)
- Pictures:
  - http://www.google.com/intl/de/about/datacenters/gallery/index.html#/

Changes in size and….

Source: http://cseweb.ucsd.edu/classes/fa12/cse291-c/talks/SCC-80-core-cern.pdf

First TeraScale* computer: 1997



Intel's ASCI Option Red

Intel's ASCI Red Supercomputer

9000 CPUs

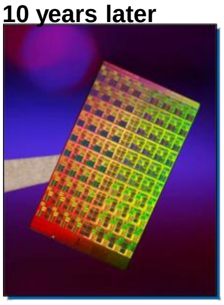
one megawatt of electricity.

1600 square feet of floor space.

*Double Precision TFLOPS running MP-Linpack

First TeraScale chip: 2007

10 years later



Intel's 80 core teraScale Chip

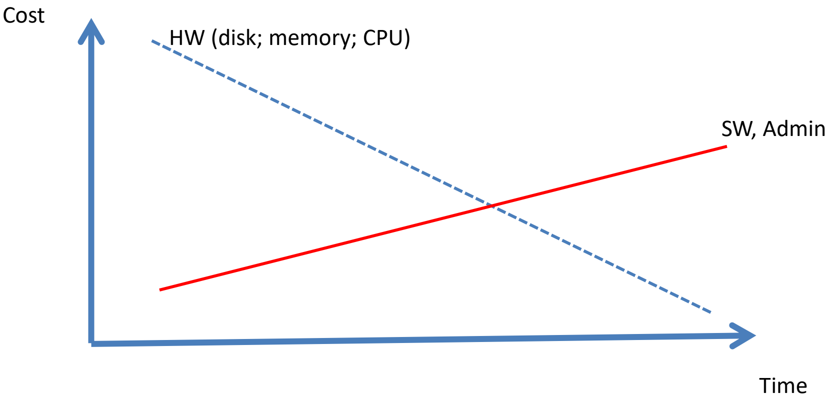
1 CPU

97 watt

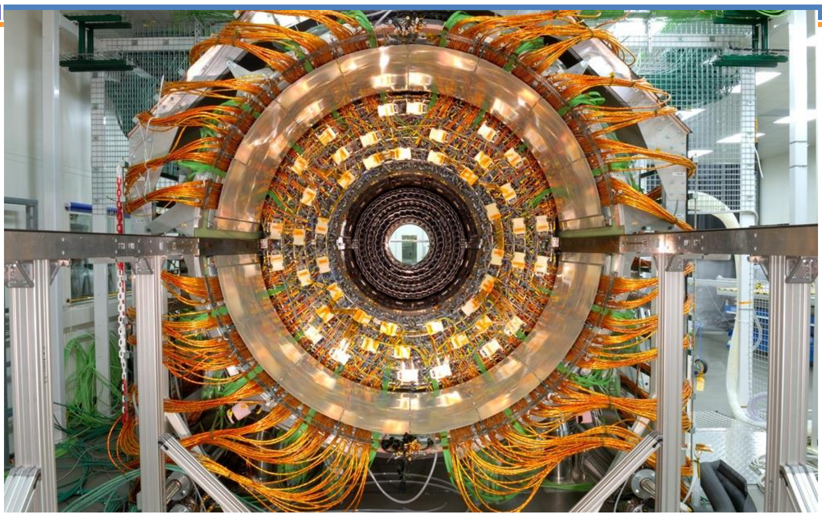
275 mm2

Single Precision TFLOPS running stencil

Cost of HW, SW, Admin



Large data – how to handle?



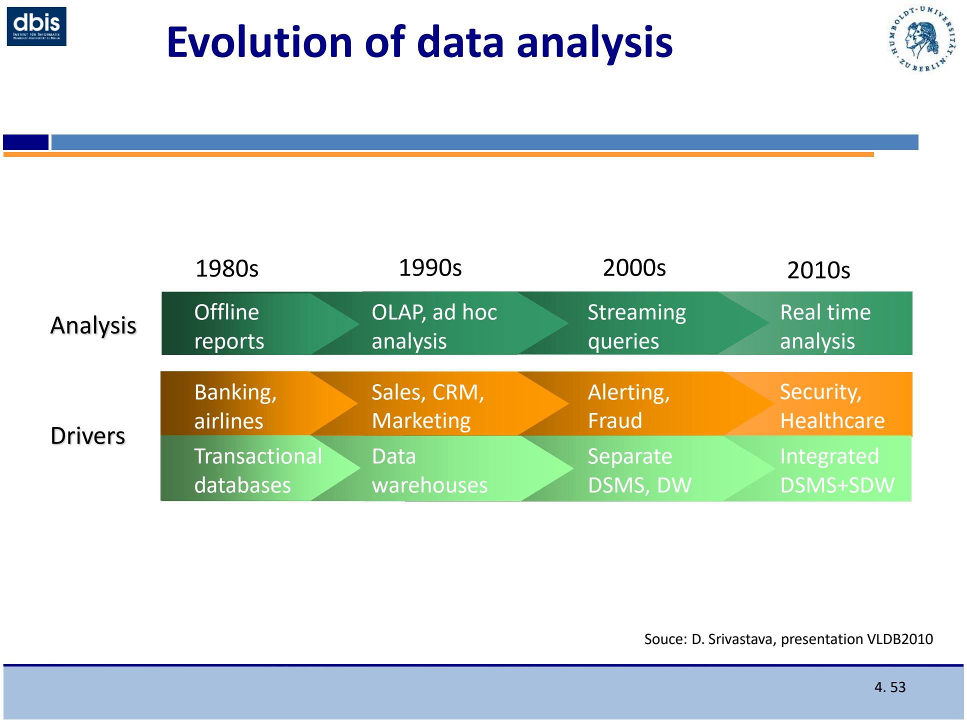
Astronomy - Skyserver

- Large volumes of data:
  - sss
- Public access:
  - http://skyserver.sdss.org/public/en/
  - Browsing the schema: http://cas.sdss.org/dr5/en/help/browser/browser.asp
- Description of project
  - http://www.sdss.org/

A few examples…

- Google processes **20 P(eta)Bytes** per day (2008)
- “All words ever spoken by a human”: ~ 5 E(xa)Bytes
- National Oceanic and Atmospheric Administration (USA): about **1 P(eta)Bytes** climate data (2007)
- CERN’s LHC generates **15 PBytes** per year (2008)

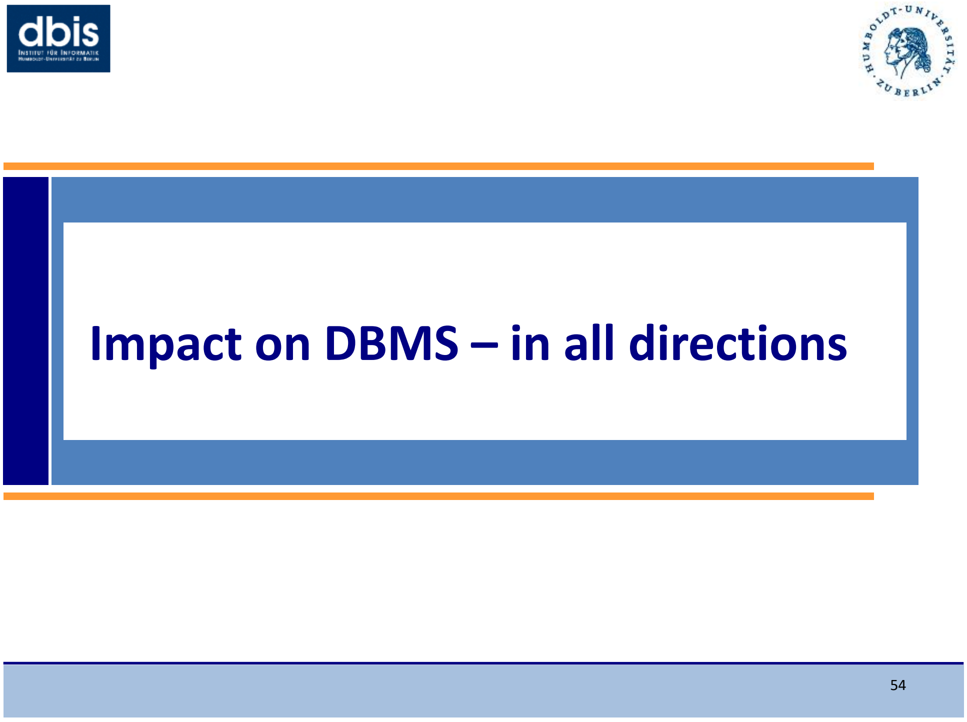
640K should be enough main memory!



Evolution of data analysis

	1980s	1990s	2000s	2010s
Analysis	Offline reports	OLAP, ad hoc analysis	Streaming queries	Real time analysis
Drivers	Banking, airlines	Sales, CRM, Marketing	Alerting, Fraud	Security, Healthcare
	Transactional databases	Data warehouses	Separate DSMS, DW	Integrated DSMS+SDW

Souce: D. Srivastava, presentation VLDB2010



Impact on DBMS – in all directions

Example for data size

- Number of US citizens: $3*10^8$

⇒ # of phone calls per citizen per day: 10

  - ⇒ $3*10^9$ phone calls per day total
  - ⇒ ~ 10^{12} phone calls per year total

⇒ 100 Bytes/phone call for recording

  - ⇒ 10^{14} Bytes per year = 100 TB per year

- Fits in main memory

What does this mean for DBMS? (3)

- □ Multicore CPUs
  - □ Main memory – cache: large gap!
    - □ Will not close up soon!
    - □ How to reduce/contain the problem?
  - □ Fine grain parallelism
    - □ How to program? – not a DB issue

      “ This rewriting […of programs…] can be done in C rather than in assembly language, using intrinsics provided in Intel's i.cc compiler.”

What does this mean for DBMS? (4)

- Unlimited # of Nodes
  - Allocate CPU nodes like main memory
  - How? On what kind of tasks?
    - Don’t save – “WASTE”!!
      - … on computations you could not have done in the past because of cost/overhead!
- Main memory is (almost) infinite (Terabytes)
  - Data always stays in MM once it’s loaded
- Cannot (Should not) admin DBMS:
  - DBMS: adaptable/self organizing
    - @ execution time
    - On all levels

Emerging HW Platform - CPU Farms

CPU Farms - Characteristics

- 1000s of („pizza“) boxes - main characteristics
  - Shared nothing
  - Data parallelism - partitioning
- Basic components & architecture
  - One or more CPUs (with many cores)
  - Local main memory
  - High speed communication adapter
  - (local disc)

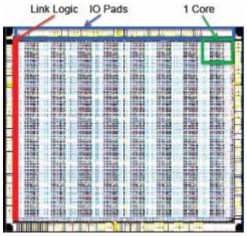
CPU Farms – Emerging Concepts

- Impact on data processing (large volumes)
  - Large volumes: Petabytes
- Potential Customers:
  - Only Google/Yahoo??
  - Sharing is necessary
- Emerging new concepts
  - Cloud computing
  - Map/Reduce compute paradigm (Architecture??)

Emerging HW Platform - Multicore

Multicore - Characteristics

- 10’s of cores - main characteristics
  - Shared nothing
  - Synchronization necessary
  - Data partitioning & data sharing
- Basic components & architecture
  - N cores in one CPUs
  - Caches (Cache hierarchy)
  - Access to local main memory

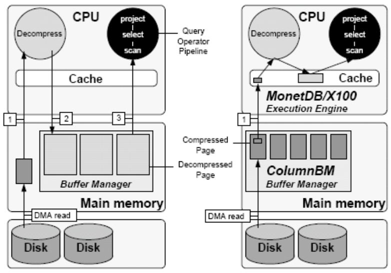


Multicore – Characteristics (2)

- Impact on data processing (large volumes)
  - Large volumes: Terabytes (??)
  - CPU intensive computation
- Potential Customers/Applications:
  - Simulations/Analysis
  - Sharing is necessary
  - ???
- Emerging new concepts
  - Main memory DBMS
  - More !!! necessary

Gap main memory & (LL) cache

- Project Monet (CWI, Amsterdam)
  - Get more “needed” data into cache
    - Column wise storage and processing (Streaming!)
    - Compressing/decompressing data



Source: Zukowski, M., Heman, S., Nes, N., & Boncz, P. (2005). *Super-scalar RAM-CPU cache compression.* Res. Rep. CWI, Amsterdam

Questions??

Google – server farms

- Movie:
  - http://www.cbsnews.com/video/watch/?id=50133304n

    (from http://tech.slashdot.org/comments.pl?sid=3191691&cid=41680953)
- Pictures:
  - http://www.google.com/intl/de/about/datacenters/gallery/index.html#/