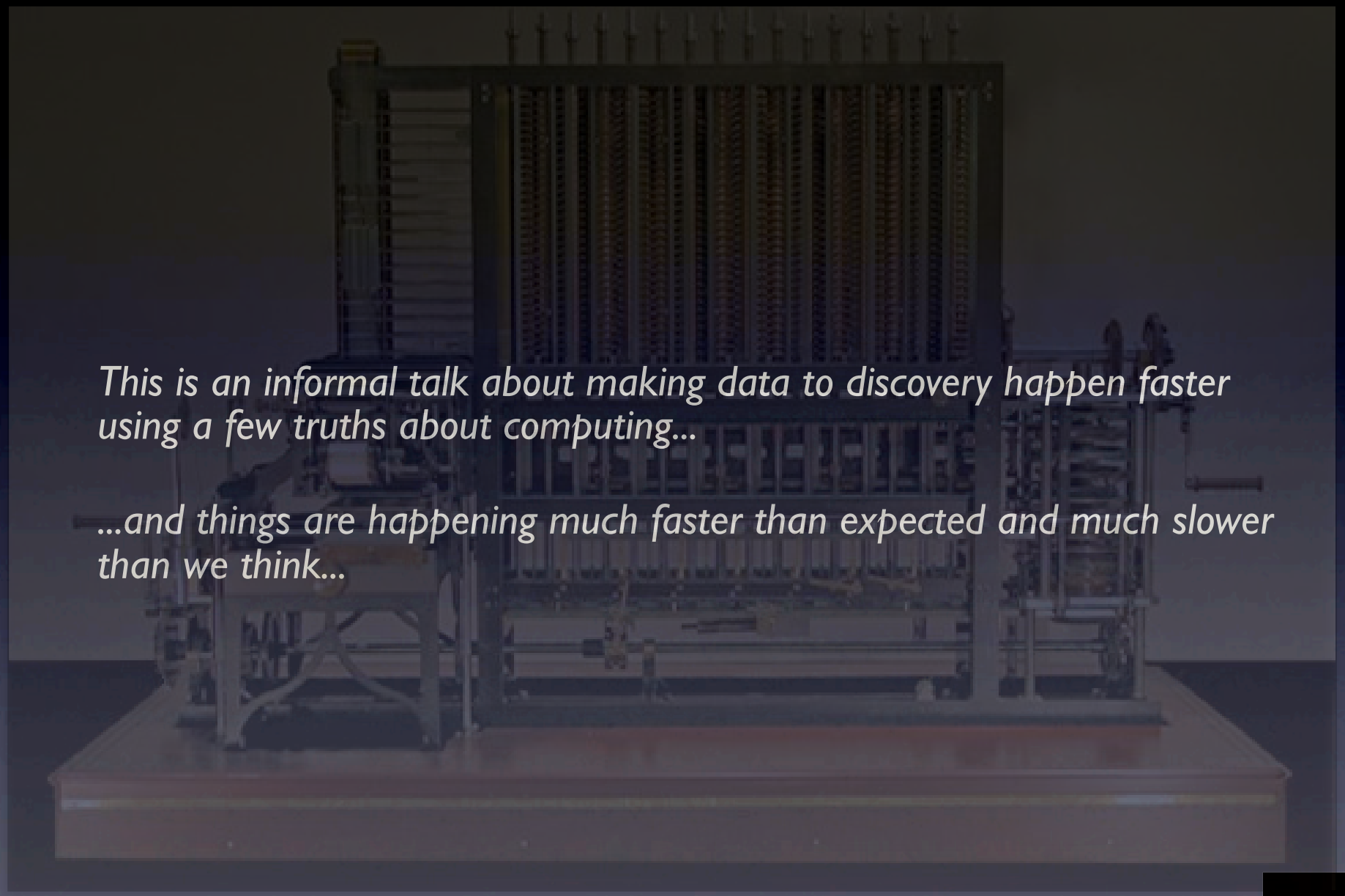
A (Hypothetical) Data to Discovery Engine

Mark Stalzer
Center for Advanced Computing Research
California Institute of Technology
stalzer@caltech.edu

AstroInformatics2010
June 16, 2010



This is an informal talk about making data to discovery happen faster using a few truths about computing...

...and things are happening much faster than expected and much slower than we think...

Source: Computer History Museum, Mt. View, CA

CACR

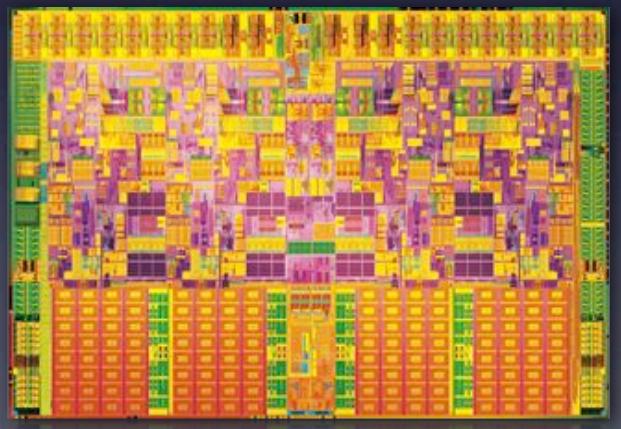
Truths

- It must be COTS (cheaply manufacturable IP)
  - Cray-1 was COTS: 6 dual in-out ECL gates per chip
  - Drivers: missile guidance and virtual missile guidance
- Advanced computing systems are all about power and packaging
  - Batteries and 100 MW power plants are expensive
  - Apple iPad and Cray-1
- Advanced computers are hard to program
  - But they can be easy to use
  - It only takes a few good abstractions

CACR

CACR & The Intel Touchstone Delta: World's Fastest Computer in 1991 (30 Gflops)

Nehalem 2009



CACR

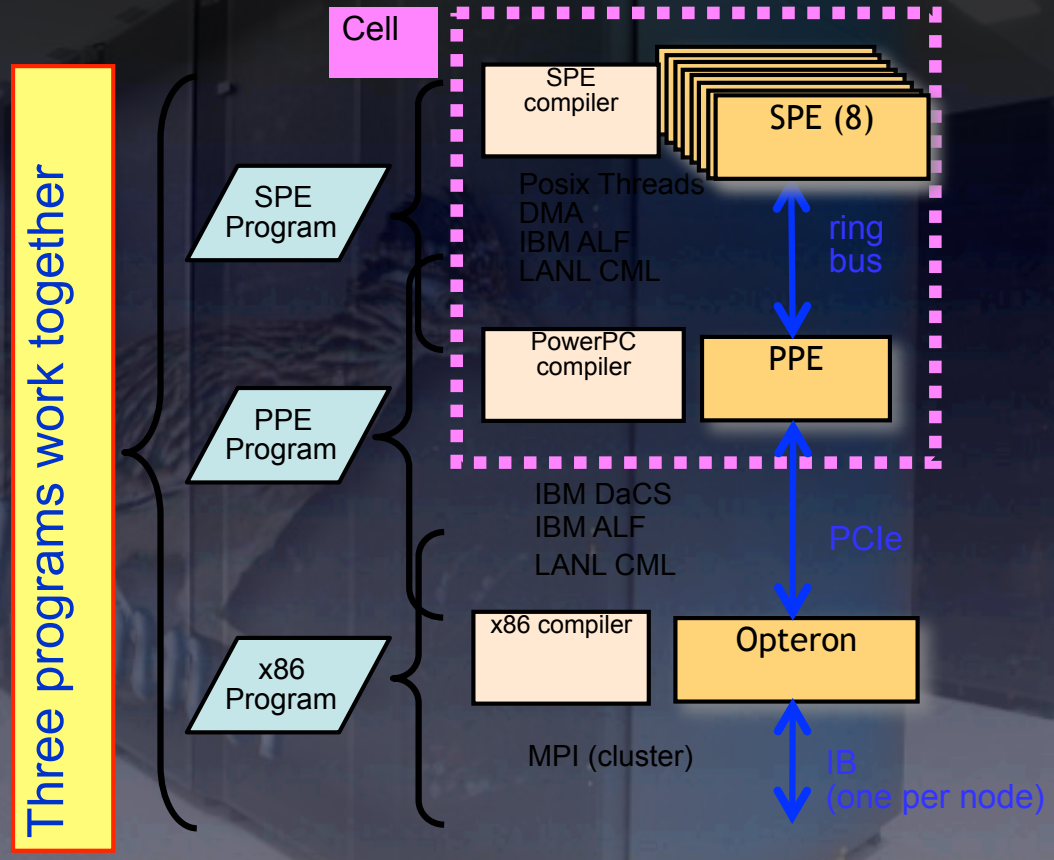
Top 10 Supercomputers

Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	Oak Ridge National Laboratory United States	Jaguar - Cray XT5-HE Opteron Six Core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.60
2	National Supercomputing Centre in Shenzhen (NSCS) China	Nebulae - Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU / 2010 Dawning	120640	1271.00	2984.30	
3	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband / 2009 IBM	122400	1042.00	1375.78	2345.50
4	National Institute for Computational Sciences/University of Tennessee United States	Kraken XT5 - Cray XT5-HE Opteron Six Core 2.6 GHz / 2009 Cray Inc.	98928	831.70	1028.85	
5	Forschungszentrum Juelich (FZJ) Germany	JUGENE - Blue Gene/P Solution / 2009 IBM	294912	825.50	1002.70	2268.00
6	NASA/Ames Research Center/NAS United States	Pleiades - SGI Altix ICE 8200EX/8400EX, Xeon HT QC 3.0/Xeon Westmere 2.93 Ghz, Infiniband / 2010 SGI	81920	772.70	973.29	3096.00
7	National SuperComputer Center in Tianjin/NUDT China	Tianhe-1 - NUDT TH-1 Cluster, Xeon E5540/E5450, ATI Radeon HD 4870 2, Infiniband / 2009 NUDT	71680	563.10	1206.19	
8	DOE/NNSA/LLNL United States	BlueGene/L - eServer Blue Gene Solution / 2007 IBM	212992	478.20	596.38	2329.60
9	Argonne National Laboratory United States	Intrepid - Blue Gene/P Solution / 2007 IBM	163840	458.61	557.06	1260.00
10	Sandia National Laboratories / National Renewable Energy Laboratory United States	Red Sky - Sun Blade x6275, Xeon X55xx 2.93 Ghz, Infiniband / 2010 Sun Microsystems	42440	433.50	497.40	

Source: top500.org 6/2010

CACR

Roadrunner



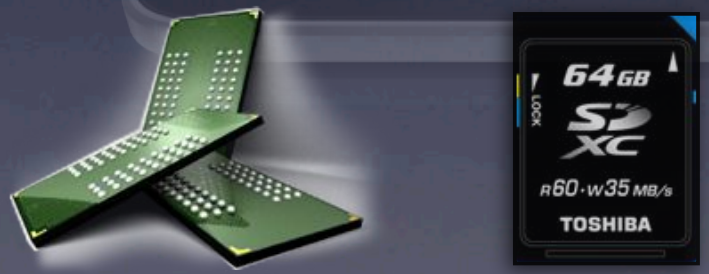
12,240 Cells each with an Opteron core
High power efficiency (acceleration)
Low cross section (reliability)

Source: M. Wingate, LANL

CACR

Portable Electronic Devices: Apple A4 to C1

- A4 guess?
  - SoC/PoP
  - ARM/GPU/USB 2.0/Flash cntrl.
  - 256 MB
  - 4 Gflops? at 1 W
  - 64 GB NAND Flash



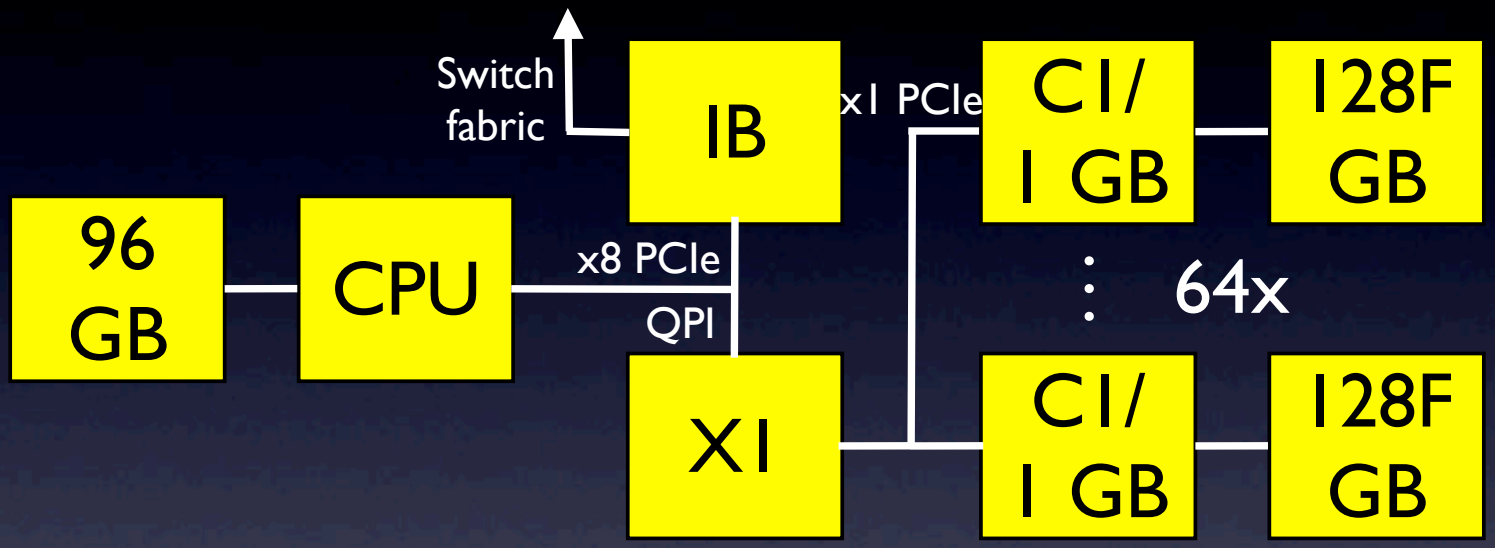
- C1
  - Proc./Accel./x1 PCIe/Flash RAID
  - 1 GB LDDR2 (64 bit wide) PoP
  - 40 Gflops at 6 W + 2 W (1 Ghz)
  - 128 GB NAND Flash RAID
  - All existing IP; need <1 year

Sources: Apple, Wikipedia, Micron, Toshiba

CACR

What do you get when you cross Roadrunner with iPads?

The Engine's Cylinders: Flashblades



- "X1" is an FPGA switch for C1 array x1 PCIe links & QPI to CPU
- The CPU orchestrates *abstractions*; to the CPU the array looks like:
  - A 6.14 TB, 25 GB/s (burst), 50 us, || disk (file system, database)
  - A 2.56 TFlops accelerator (OpenCL with DB access)
- This all fits on a standard blade (2 sides) and uses commodity IP

CACR

Blade & Server Specifications

- Flashblade
  - C1 array (64x 40 Gflops/1 GB): 2.56 Tflops/64 GB
  - Flash array (64x 128 GB 75% RAID): 6.14 TB useable
  - X I/O bandwidth: 32 GB/s in C1 array & 25 GB/s to CPU
  - Power (64x @ 8 W + 100 W): ~600 W
  - Estimated cost per blade: $25K (not including C1&blade NRE)
- Blade server (e.g. IBM BladeCenter/E 14x at 7U)
  - C1 arrays: ~36 Tflops & ~900 GB
  - Flash arrays: ~86 TB
  - Network (14x IB @ 40 Gb/s): 70 GB/s
  - Power: 8400 W for blades (BC/E max is 9300 W)
  - Estimated cost per server: $400K = $350K+$50K system

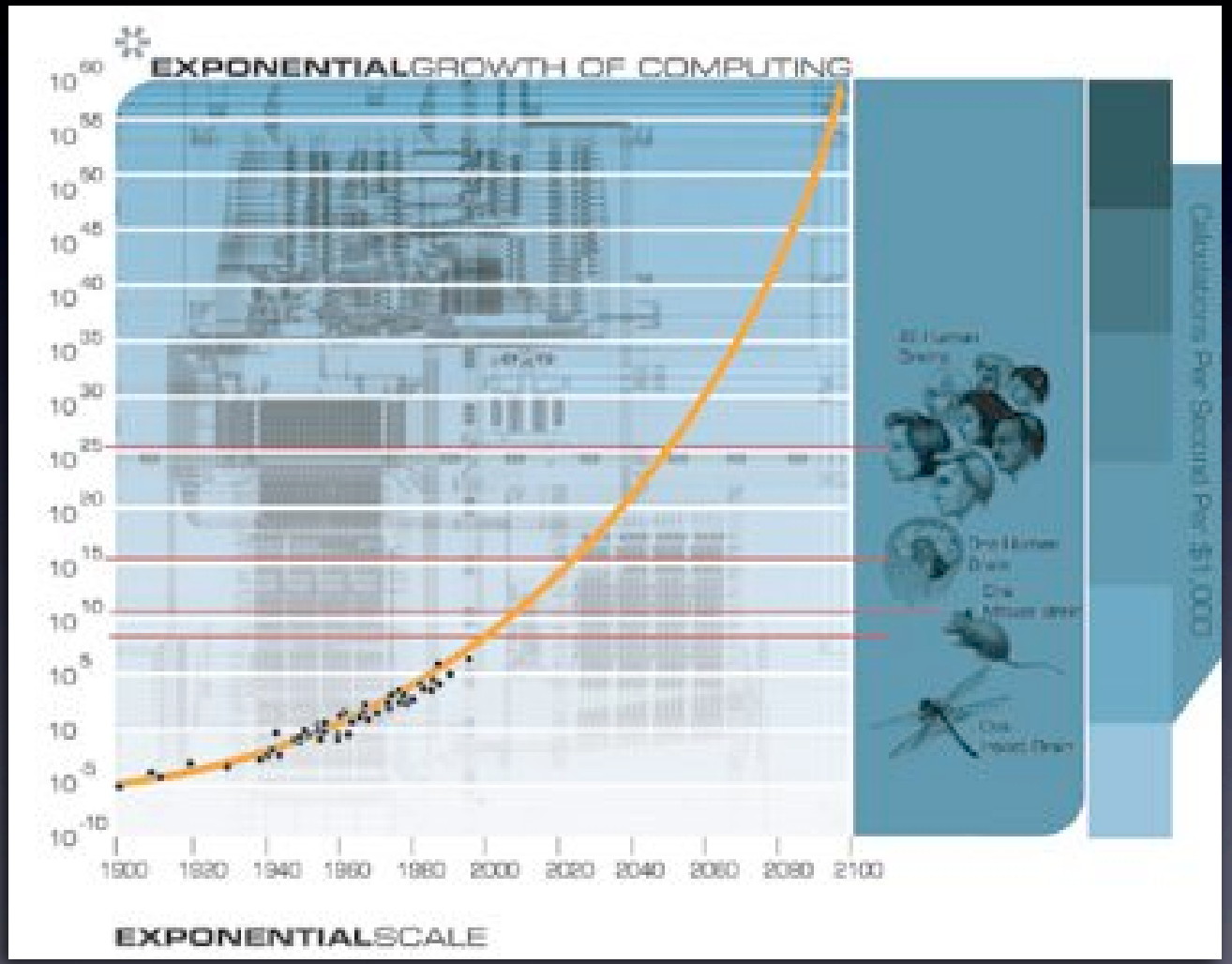
CACR

Implications for Data to Discovery

- ***HUGE*** data processing capability: a single server can read (and "process") its entire contents in about 9 minutes (240 MB/s)
  - A same size disk array would take 75x to read (2 TB disks)
  - 100x faster at random access (50 us vs. 5 ms)
  - The accelerator balances read bandwidth
- ***This is qualitatively new, what could we do with it?***
- Want applications with large #reads to #writes
  - One rack (0.5 PB) could handle LSST processing for a year?
  - Good for LHC/CMS data analysis (comparing data to Monte-Carlo simulations of the standard model)
  - Analytics, search, intelligence, video processing, tomography, ...
  - Metadata server for massive archival disk arrays

CACR

It Only Takes 10^{16} Ops (?)



Source: R. Kurzweil, The Law of Accelerating Returns, 3/2001, The Singularity is Near, pg. 124, 2005

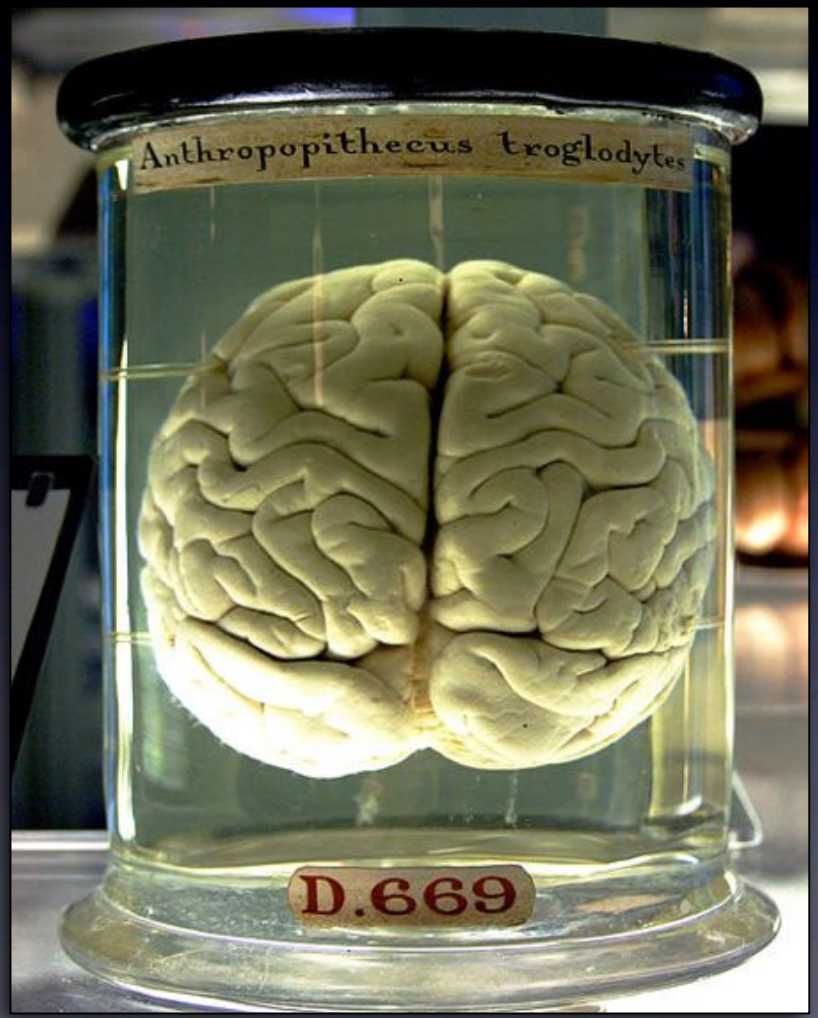
CACR

A 10^{16} Flops Engine

- Quantity: Need about 300 servers (50 racks)
  - Big IB switch fabric too (84 IB ports/rack)
  - Volume (need lots of air): 10,000 ft^3
- Flash "never forget" memory (mirror in system): ~13 PB
  - SDSC runs 18 PB of tape storage (1,000's of scientific data sets)
  - Constraint: no more than ~1,000 writes/chip/day
  - Data Ingestion and Reflective Analysis by engine
  - Checkpoints in about 10 s
- About 10 Pflops at 3 MW (does not include fabric or cooling)
- Cost unknown due to rapidly dropping flash costs, repackaging, and other economies of scale
  - Guess about $100M acquisition (2012)

CACR

Comparison to Another Data to Discovery Engine

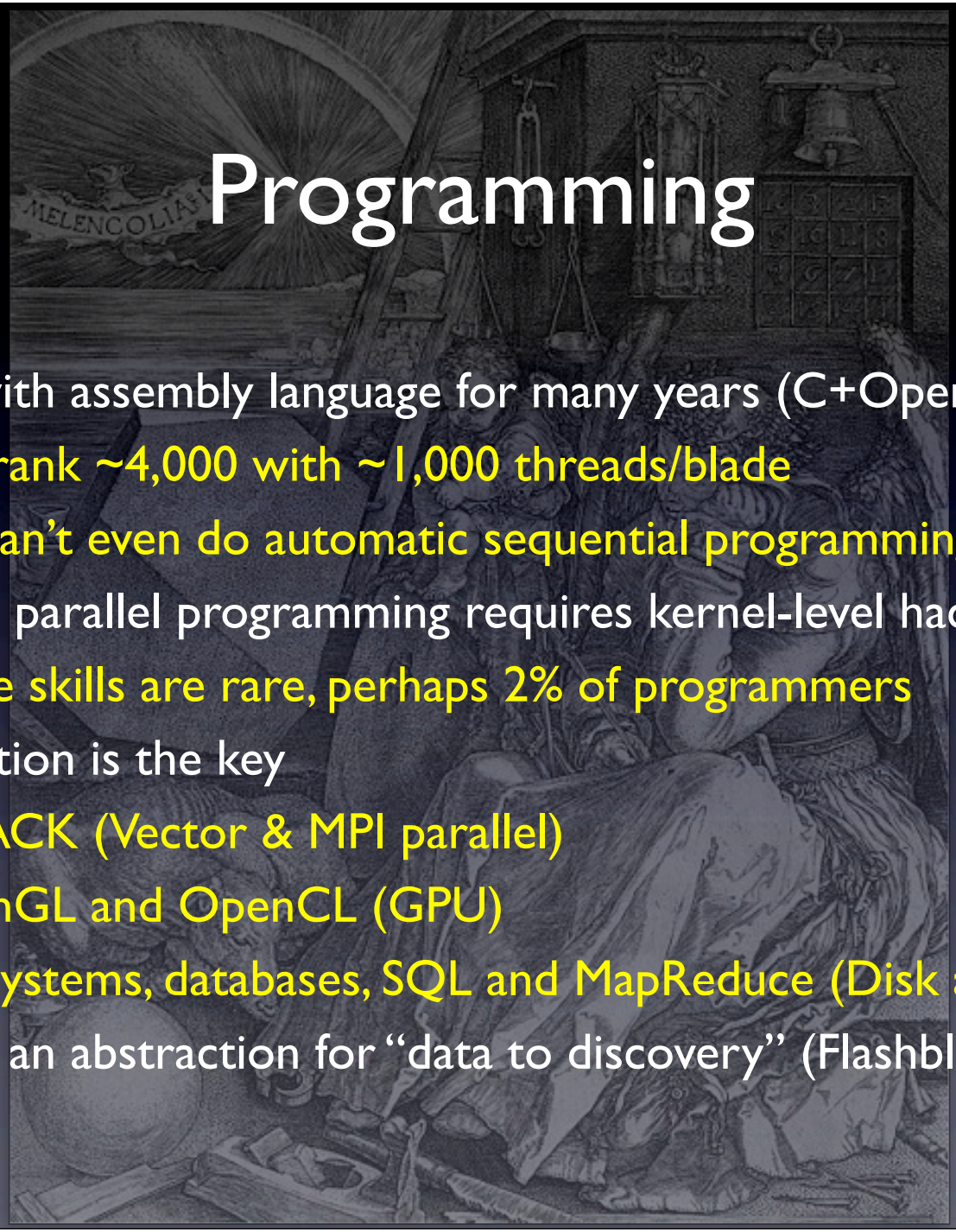


- Operations: 10 Pops (1x)
- Memory: 1 PB (0.1x & forgets)
- Bandwidth: 1 PB/s? (1x - 10x - 50x)
- Volume: 0.25 ft^3 (40,000x)
- Power: 25 W (120,000x)!
- ***Where's the algorithm?***

Sources: Brain picture: Wikipedia; Memory: R. Kurzweil, pg 127; Bandwidth: author

CACR

Programming



- Stuck with assembly language for many years (C+OpenCL+MPI)
  - MPI rank ~4,000 with ~1,000 threads/blade
  - We can't even do automatic sequential programming
- Scalable parallel programming requires kernel-level hacking skills
  - These skills are rare, perhaps 2% of programmers
- Abstraction is the key
  - LAPACK (Vector & MPI parallel)
  - OpenGL and OpenCL (GPU)
  - File systems, databases, SQL and MapReduce (Disk arrays)
- What is an abstraction for "data to discovery" (Flashblades)?

Source: A. Dürer, Melencolia I

CACR

Some Characteristics of Good Abstractions

- The abstraction must be scientifically useful
- Specify *what* not *how*
  - E.g. factor a matrix or render an image
- The *what* should be implicitly parallel in some way
  - Map over matrix elements, polygons, pixels, rows, columns, ...
  - Implementations provide fault tolerance
- *The fundamental operations must lead to global error bounds and convergence rates (algorithms)*
  - This is crucial because the algorithms might do 10^{21} ops
- Parameterized (consistent with the above)
  - Error bounds, data types, realism level, ...

Conclusions

- We are not stuck with "clusters": COTS is also IP not just Fry's
- Trans petascale data computing systems can be built *now*
  - Hardware and software are not necessarily barriers
  - It will require investment
  - *Two orders of magnitude step-up possible for data intensive systems*

- ***I think a challenge for this community is finding good abstractions for "data to discovery"***
  - Implementing the abstractions might require algorithm research
  - This is the key interplay between Science & Computer Science
  - Efficient implementations of good abstractions will happen

CACR