The Future of Large-Scale Computing

Dan Reed
reed@microsoft.com
Multicore and Scalable Computing Strategist
Managing Director, Data Center Futures

Microsoft

Presentation Outline

- Innovation lessons
  - Disruptive technologies
  - Sapir-Whorf and the future
- Multicore futures
  - Homogeneity/heterogeneity
- The data deluge
  - Insight from data
- Cloud economics
  - Spending Moore's dividend
- Cloud data centers and exascale
  - Designing for the future

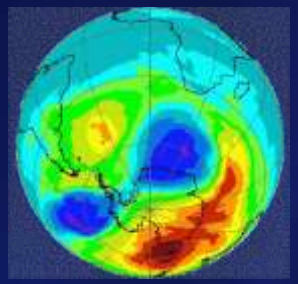
Fortitudine Vincimus!

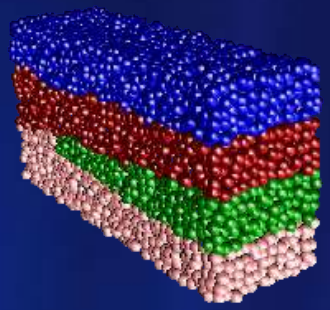
Microsoft

Science 2020

"In the last two decades advances in computing technology, from processing speed to network capacity and the Internet, have revolutionized the way scientists work.

From sequencing genomes to monitoring the Earth's climate, many recent scientific advances would not have been possible without a parallel increase in computing power - and with revolutionary technologies such as the quantum computer edging towards reality, *what will the relationship between computing and science bring us over the next 15 years?"*



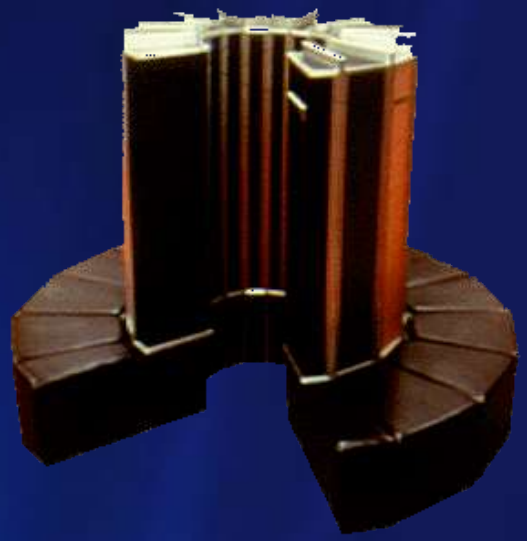


http://research.microsoft.com/towards2020science

Microsoft

Sapir–Whorf: Context and Research

- Sapir–Whorf Hypothesis (SWH)
  - Language influences the habitual thought of its speakers
- Scientific computing analog
  - Available systems shape research agendas
- Consider some past examples
  - Cray-1 and vector computing
  - VAX 11/780 and UNIX
  - Workstations and Ethernet
  - PCs and web
  - Inexpensive clusters and Grids
- Today's examples
  - multicore, sensors, clouds and services …
- What lessons can we draw?



Microsoft

Today's Truisms (2008)

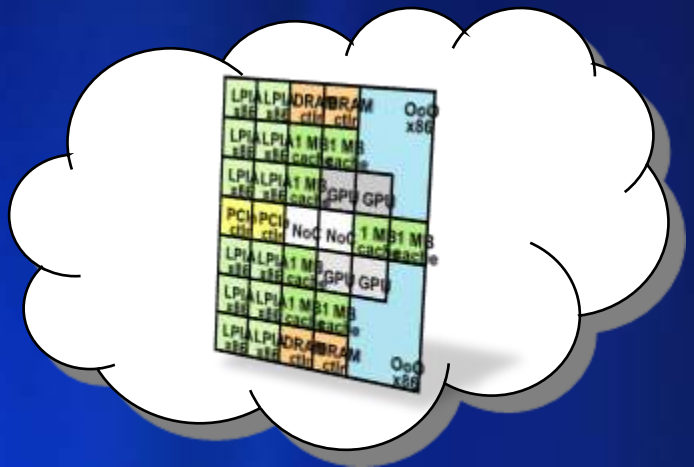
- Bulk computing is almost free
  - … but software and power are not
- Inexpensive sensors are ubiquitous
  - … but scientific data fusion remains difficult
- Moving lots of data is {still} hard
  - … because we're missing trans-terabit/second networks
- People are @#$(& expensive!
  - … and robust software remains extremely labor intensive
- Scientific challenges are complex
  - … and social engineering is not our forte
- Our political/technical approaches must change
  - … or we risk solving irrelevant problems

The Pendulum Swings

- Mainframes
  - Remote job entry (RJE)
- Minicomputers
  - Distributed processing
- Workstations
  - Local area networks, X and DCE
- PCs and Internet
  - Desktop clients and browsers
- Grids and clusters
  - Linux and web services
- Clouds and ManyCore
  - Consumer parallelism
  - Software as a service

The Big Sucking Sound …

- Not the financial markets …
- It's that other sound …
- The pull of commercial economics
  - Manycore processors and accelerators
  - Software as a service and cloud computing



Microsoft

Moore's "Law" and the Dividend

The experts look ahead

Cramming more components onto integrated circuits

With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single silicon chip

By Gordon E. Moore

Director, Research and Development Laboratories, Fairchild Semiconductor division of Fairchild Camera and Instrument Corp.

The future of integrated electronics is the future of electronics itself. The advantages of integration will bring about a proliferation of electronics, pushing this science into many new areas.

Integrated circuits will lead to such wonders as home computers—or at least terminals connected to a central computer—automatic controls for automobiles, and personal portable communications equipment. The electronic wristwatch needs only a display to be feasible today.

But the biggest potential lies in the production of large systems. In telephone communications, integrated circuits in digital filters will separate channels on multiplex equipment. Integrated circuits will also switch telephone circuits and perform data processing.

Computers will be more powerful, and will be organized in completely different ways. For example, memories built of integrated electronics may be distributed throughout the machine instead of being concentrated in a central unit. In addition, the improved reliability made possible by integrated circuits will allow the construction of larger processing units. Machines similar to those in existence today will be built at lower costs and with faster turn-around.

Present and future

By integrated electronics, I mean all the various technologies which are referred to as microelectronics today as well as any additional ones that result in electronics functions supplied to the user as irreducible units. These technologies were first investigated in the late 1950's. The object was to miniaturize electronics equipment to include increasingly complex electronic functions in limited space with minimum weight. Several approaches evolved, including microassembly techniques for individual components, thin-film structures and semiconductor integrated circuits.

Each approach evolved rapidly and converged so that each borrowed techniques from another. Many researchers believe the way of the future to be a combination of the various approaches.

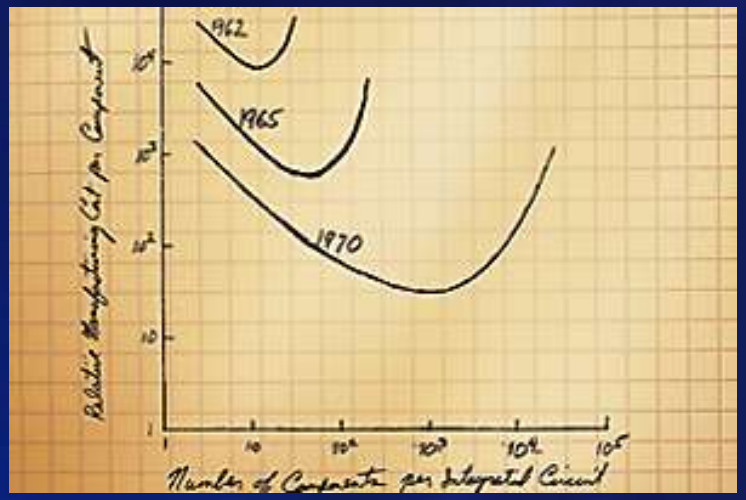
The advocates of semiconductor integrated circuitry are already using the improved characteristics of thin-film resistors by applying such films directly to an active semiconductor substrate. Those advocating a technology based upon films are developing sophisticated techniques for the attachment of active semiconductor devices to the passive film arrays.

Both approaches have worked well and are being used in equipment today.

The author

Dr. Gordon E. Moore is one of the new breed of electronic engineers, schooled in the physical sciences rather than in electronics. He earned a B.S. degree in chemistry from the University of California and a Ph.D. degree in physical chemistry from the California Institute of Technology. He was one of the founders of Fairchild Semiconductor and has been director of the research and development laboratories since 1959.

Electronics, Volume 38, Number 8, April 19, 1965



Microsoft

Expectations Evolve

State of the Art, ~1981

IBM PC
Intel 8080 @ 4.77 MHz
16-640 KB memory

Aspirational, ~1981

PERQ “3M” machine
1 MIPS, 1 MB, 1 megapixel

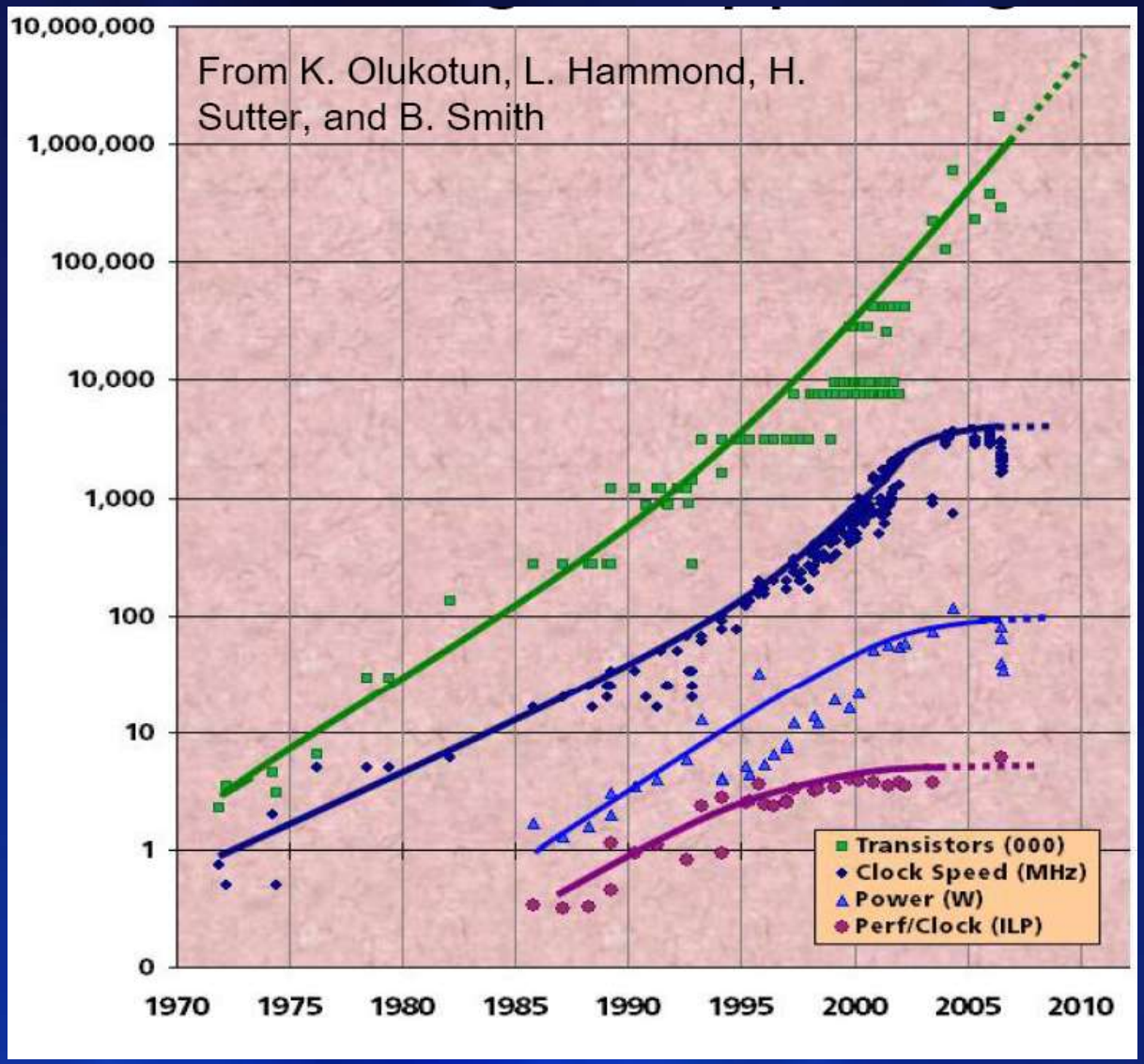
HPC Expectations Evolve

- ASCI Red (1997)
  - ~10,000 Pentium Pro
    - 200 MHz
  - 200 MF/processor
  - ~1 MW total
  - ~2500 sq ft (230 m^2)

- ASC Roadrunner (2008)
  - ~130,000 cores
    - 6948 dual core 1.8 GHz Opterons
    - 12960 3.2 GHz Cell processors
  - ~400 GF/node
  - ~3 MW total
  - ~5500 sq ft

Microsoft

I Have One Word For You: Parallelism



- Think about
  - Disks
  - FLASH
  - PCM

  and the future

Microsoft

Myhrvold's Laws, ~1997

- 1st Law
  - Software is a gas!
- 2nd Law
  - Initial growth is rapid - like gas expanding (like browser)
  - Eventually, limited by hardware (like NT)
  - Bring any processor to its knees, just before the new model is out
- 3rd Law
  - That's why people buy new hardware - economic motivator
  - That's why chips get faster at same price, instead of cheaper
  - Will continue as long as there is opportunity for new software
- 4th Law
  - It's impossible to have enough
    - New algorithms, New applications and new users
    - New notions of what is cool

Source: Jim Larus

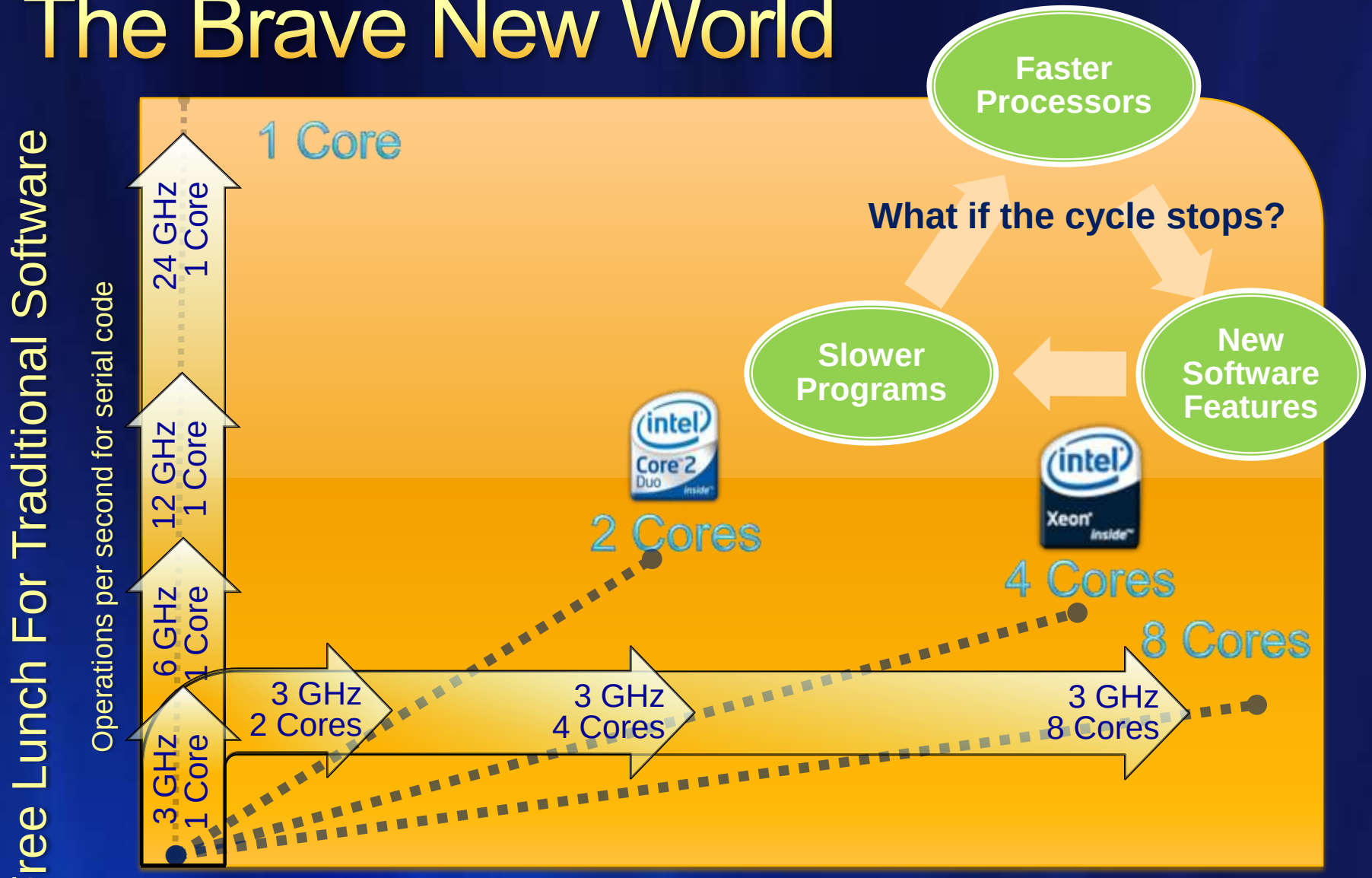
Performance Dulls Typical Behavior

- Greater performance/memory size dulls programmers'
  - Gates changed "READY" to "OK" in Altair Basic
    - Why? To save a few bytes!
- Little understanding of processor performance models
  - Who really understands cache behavior?
- Increasing reliance on compiler optimization
  - Uniformly "good" quality
  - Sometimes 10-100x off hand-written code
- Performance is not an abstraction
  - Cuts across software abstractions
  - Cannot be understood locally
    - Think locally, act globally?

Source: Jim Larus

13

The Brave New World

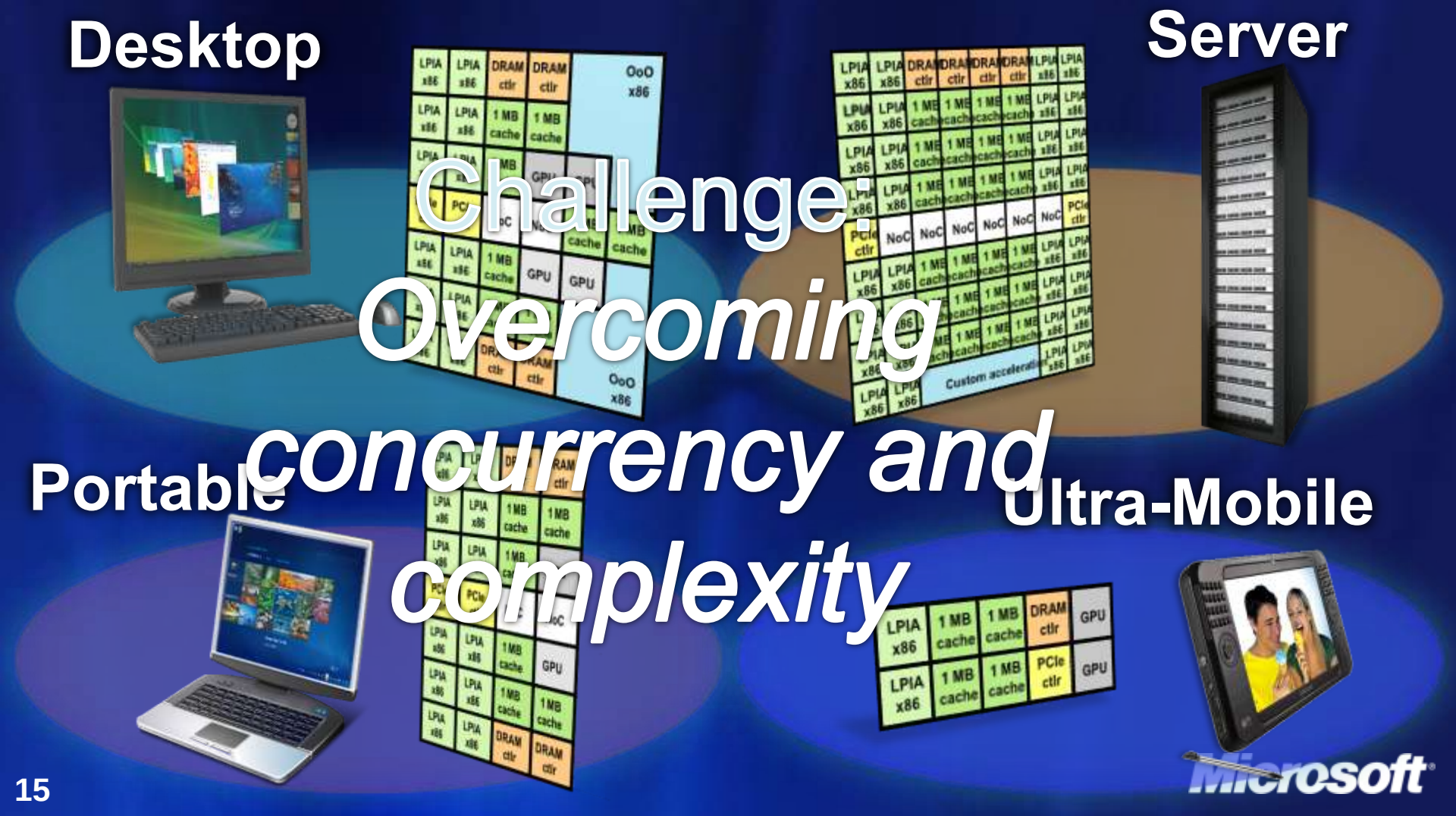


Additional operations per second if code can take advantage of concurrency

No Free Lunch For Traditional Software

(Without highly concurrent software, it won't get any faster!)

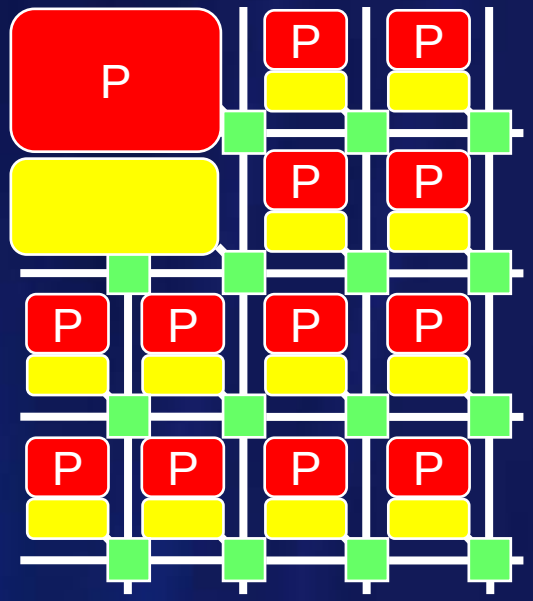
Manycore CPUs



Challenge: *Overcoming concurrency and complexity*

Core Complexity Implications

- Remember Amdahl's Law Speedup = $(S + (1-S)/N)^{-1}$
  - Some very nice recent work by Mark Hill
    - "Amdahl's Law in the Multicore Era," M. D. Hill and M. R. Marty, *IEEE Computer*, July 2008
- Many on-chip possibilities
  - PIM and mixed processes
  - Network protocol processing
  - Optical interconnect (MEMS/waveguide)
  - Crypto-processing
  - DSP/image processing and rendering
- Multicore implications
  - Symmetric or asymmetric cores
  - Legacy and new code
  - Programming heterogeneity
  - Reliability and interconnect



Microsoft

We Are Surrounded By "Opportunities"

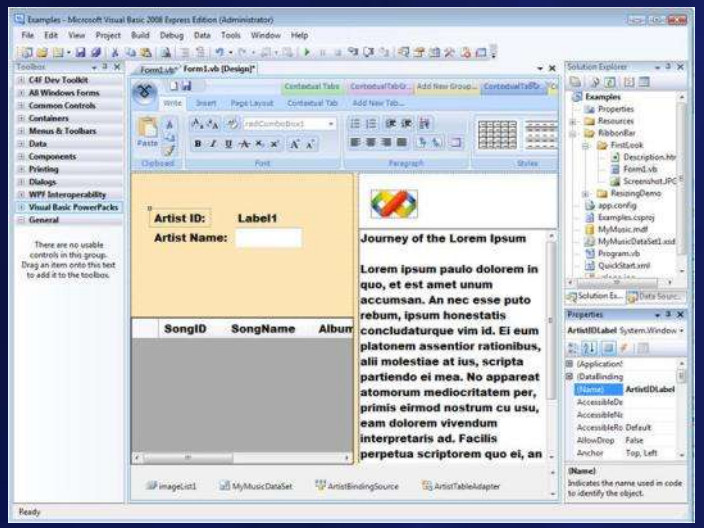
- Process variation
- Core multiplicity
- Heterogeneity (functional and performance)
- Bandwidth and latency
- Reliability and failures
- Application evolution and revolution

- Possible implications
  - Don't fight the last war – anticipate the new one
  - Embrace application adaptation during execution
    - From below (hardware) and above (user expectations)

Technical Programming Groups

- Three developer groups
  - Heroes
  - Mainstream
  - Entry/Novice
- Each with differing needs
- Petascale heroes

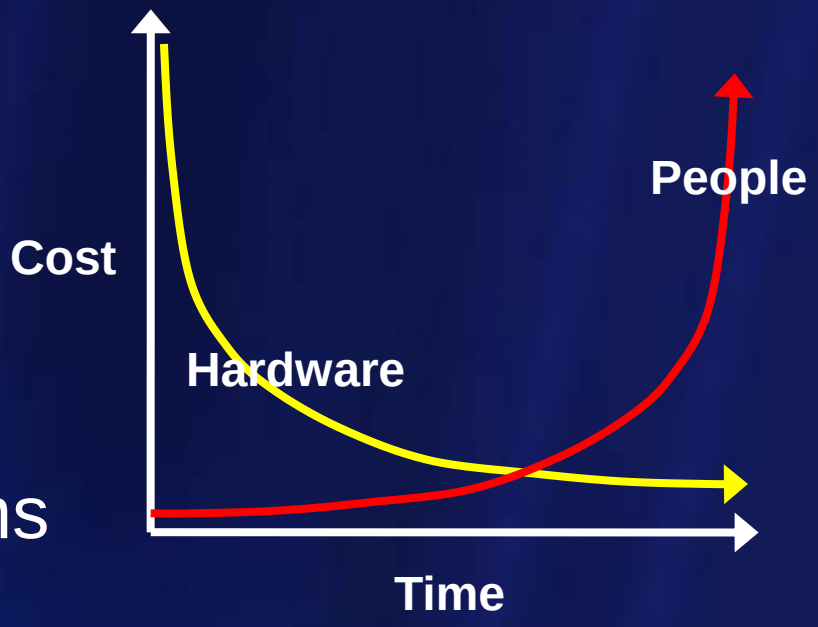
  "Neurosurgery? No problem! Hand me that sake bottle and a screwdriver."
- Mainstream
  - Savvy computational scientist
- Entry/novice
  - The typical scientist/engineer



Microsoft

Economic Divergence/Optimization

- $/teraflop-year
  - declining rapidly
- $/developer-year
  - rising rapidly
- Applications outlive systems
  - by many years



- Machine-synthesized and managed software
  - getting cheaper and more feasible …
- Two timeframes
  - compilation and execution

Microsoft

Today's HPC Environment

Clusters/ supercomputers

High speed networking

Storage

Scientists

Engineers

Financial analysts

Compilers

Specialized languages

Debuggers

Microsoft

Top 500 List (June 2008)

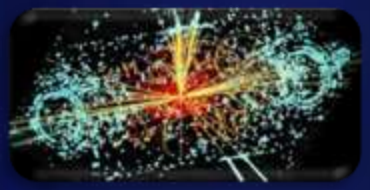
	Computer	Rmax [TF/s]	Rmax / Rpeak	Installation Site	Country	#Cores	Power [MW]	MFlops/ Watt
1	IBM / Roadrunner BladeCenter QS22/LS21	1,026	75%	DOE/NNSA/LANL	USA	122,400	2.35	437
2	IBM / BlueGene/L eServer Blue Gene Solution	478	80%	DOE/NNSA/LLNL	USA	212,992	2.33	205
3	IBM / Intrepid Blue Gene/P Solution	450	81%	DOE/OS/ANL	USA	163,840	1.26	357
4	SUN / Ranger SunBlade x6420	326	65%	NSF/TACC	USA	62,976	2.00	163
5	CRAY / Jaguar Cray XT4 QuadCore	205	79%	DOE/OS/ORNL	USA	30,976	1.58	130
6	IBM / JUGENE Blue Gene/P Solution	180	81%	Forschungszentrum Juelich (FZJ)	Germany	65,536	0.50	357
7	SGI / Encanto SGI Altix ICE 8200	133.2	77%	New Mexico Computing Applications Center	USA	14,336	0.86	155
8	HP / EKA Cluster Platform 3000 BL460c	132.8	77%	Computational Research Laboratories, TATA SONS	India	14,384	1.60	83
9	IBM / Blue Gene/P Solution	112	81%	IDRIS	France	40,960	0.32	357
10	SGI / Altix ICE 8200EX	106	86%	Total Exploration Production	France	10,240	0.44	240

Source: Jack Dongarra

The Data Explosion

Experiments

Simulations



Archives

Literature

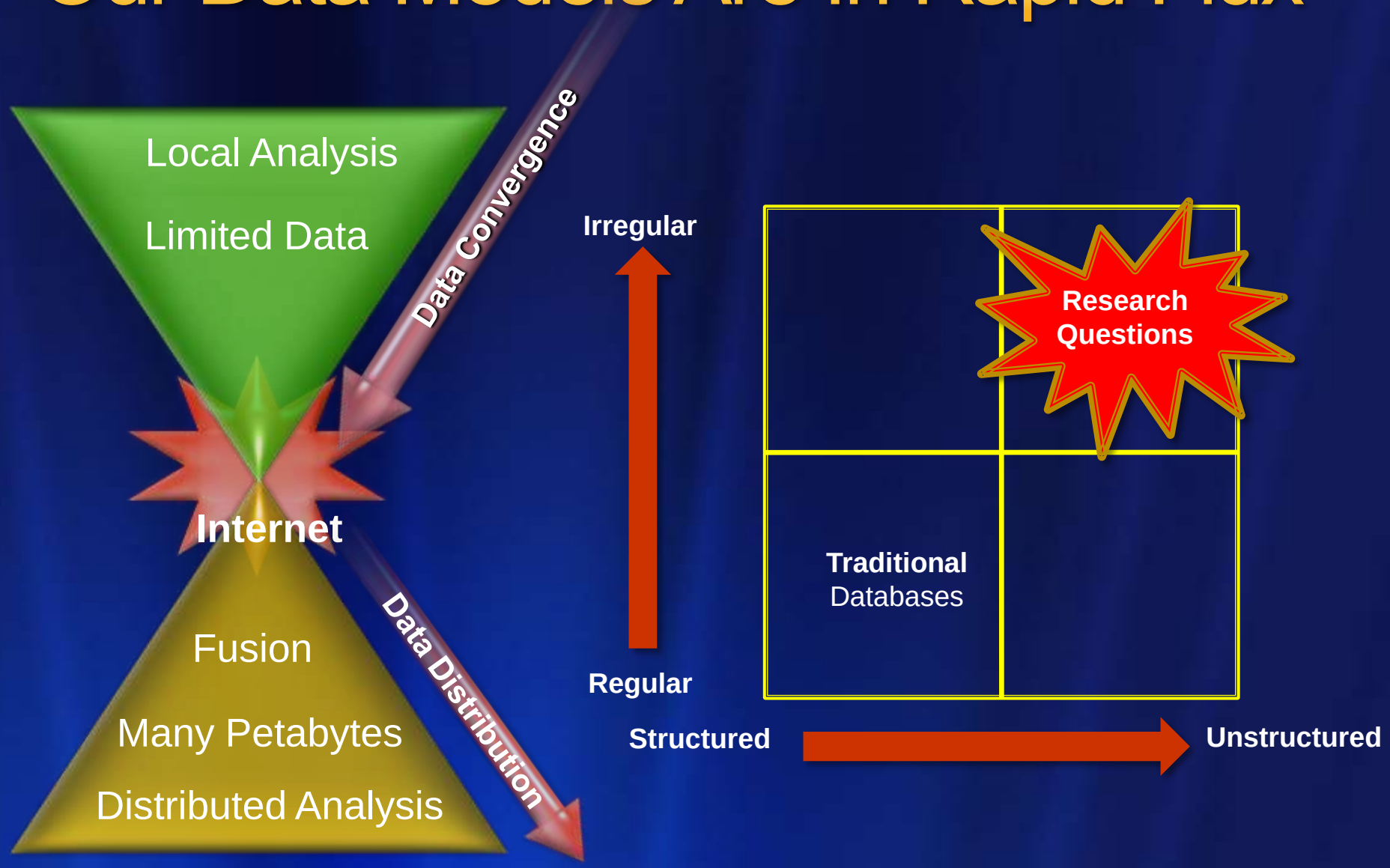
Cornell University Library

arXiv.org

<u>Petabytes</u>
Doubling every
2 years

Microsoft

Our Data Models Are In Rapid Flux

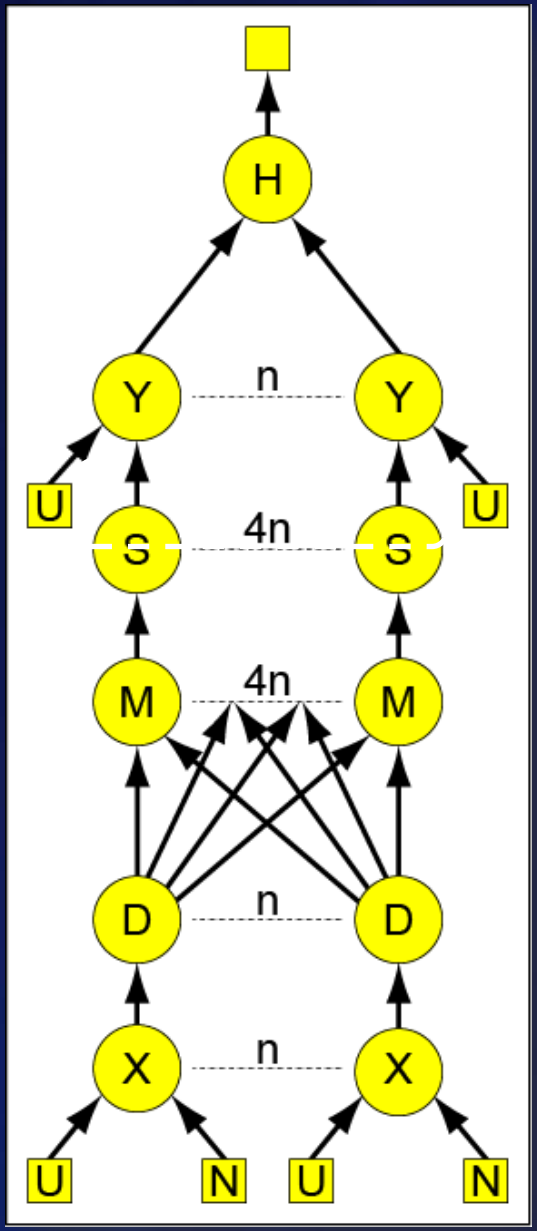


Microsoft

Dryad (MSR SVC)

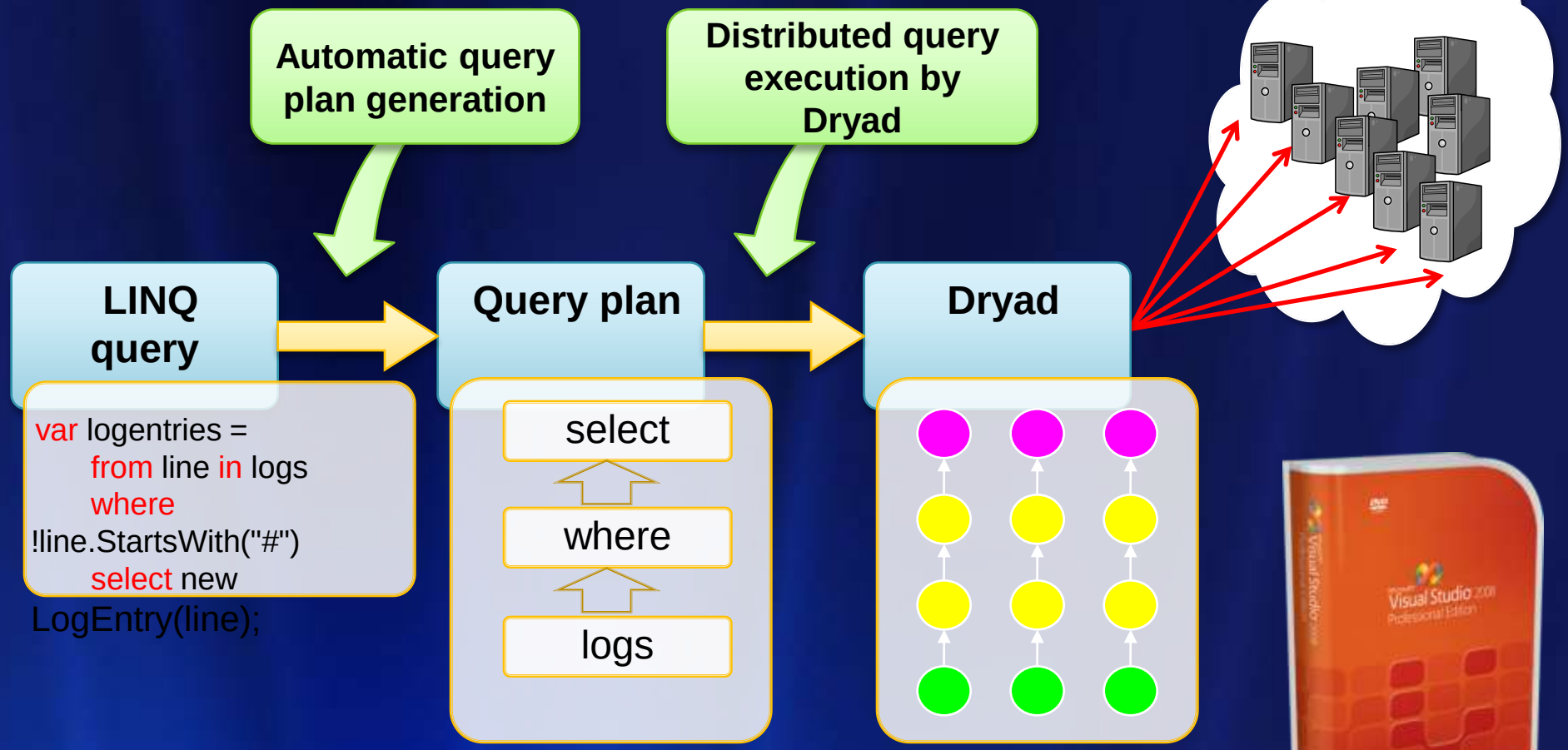
- Coarse grain data flow
  - DAG specification
    - superset of Map-Reduce
- Communication substrates
  - NTFS, TCP, shared memory
- Failure recovery
- Scheduling

- SkyServer query example
  - 3-way join to find gravitational lens



Source: Michael Isard

DryadLINQ: From LINQ to Dryad



- LINQ: .NET Language Integrated Query
  - Declarative SQL-like programming with C# and Visual Studio
  - Easy expression of data parallelism
  - Elegant and unified data model

Source: Yuan Yu *et al*

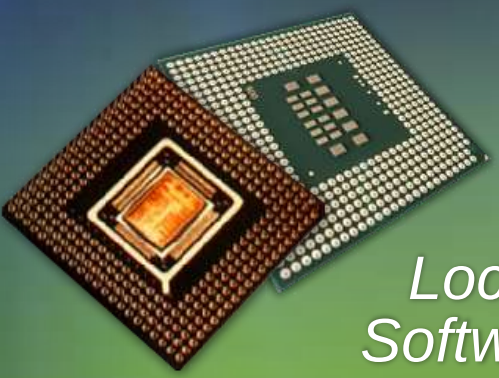
Microsoft

Insight, Not Data

What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.

Herbert Simon

Next-Generation Applications



Concurrency Spectrum

Local Software

Software + Services

Global Services

Economics Drive Change

- Moore's "Law" favored consumer commodities
  - Economics drove enormous improvements
  - Specialized processors and mainframes faltered
  - The commodity software industry was born
- Hard to compete against 50%/year improvement

- Implications
  - Consumer product space defines outcomes
    - It may not go where we hope or expect
  - Research environments track consumer trends
    - Driven by market economics

28

User Experience Waves

Can Multicore Supplant Clock Rates?

- Double the number of cores instead of speed
- No, at least without major innovation
  - Sequential code
  - Lack of parallel algorithms
  - Difficult programming
  - Few abstractions

- Parallelism will change the computing landscape
- If existing applications cannot use parallelism
  - New applications and systems will arise
    - Software plus services
    - Mobile computing

New Software Architecture

31

The Cloudy Infosphere

Microsoft

The Microsoft Services Foundation

Microsoft FlexGo

Microsoft Exchange Hosted Services

Microsoft Office Live WorkSpaces

Live Search

Microsoft Office Live Meeting

Microsoft Forefront

Microsoft Office Online

Windows Live ID

Microsoft Hosted Messaging & Collaboration

Windows Live Hotmail

Windows Live SkyDrive

Microsoft Silverlight

Sharepoint.Microsoft.com

Microsoft Office Groove 2007

Windows Live Spaces

XBOX LIVE

Microsoft HealthVault

Windows Live Search

msn

Zurich .Net Online

Office Labs

Live Search Maps

Connected Services SANDBOX
Enabling Managed Network Mashups

Plus over 150 more sites and services

msn adCenter

Microsoft game studios

Windows Live Messenger

MBS Online

Windows Live

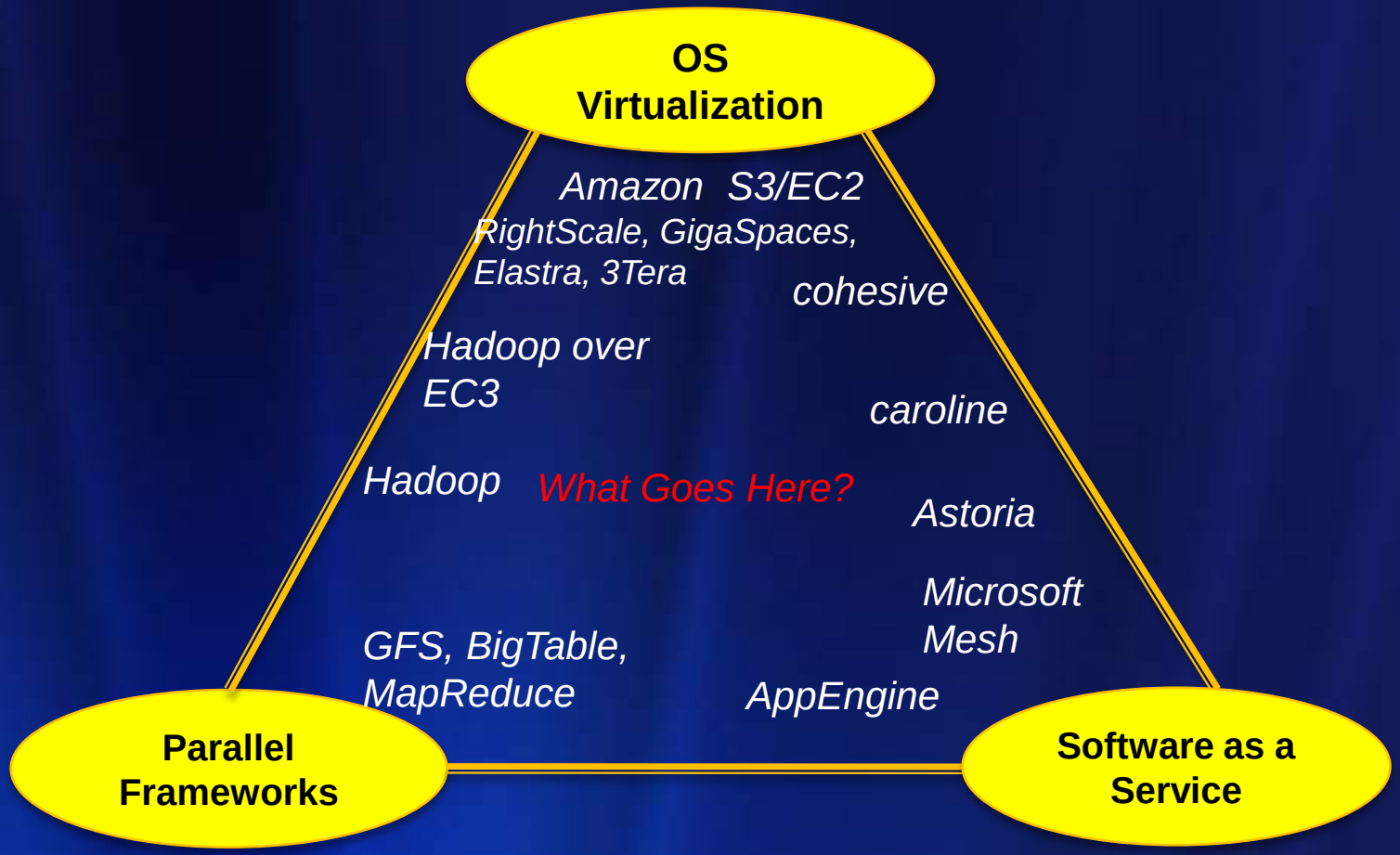
Microsoft

Embarrassingly Parallel Processing

- When applications are hosted
  - Even sequential ones are embarrassingly parallel
  - Few dependencies among users
- Moore's benefits accrue to platform owner
  - 2x processors →
    - ½ servers (+ ½ power, space, cooling …)
    - Or 2x service at the same cost
- Tradeoffs not entirely one-sided due to
  - latency, bandwidth, privacy, off-line considerations
  - capital investment, security, programming problems

34

Cloud Application Frameworks



- Spanned by three points, each defining an approach
  - Exploit parallelism, Ease deployment, Provide service access

Source: Dennis Gannon

Microsoft Clouds

We will introduce one in about ~~four~~ two weeks," We'll even have a name to give you by then. But let's just call it for the purposes of today 'Windows Cloud'."

Steve Ballmer

Microsoft Professional Developer's Conference
www.microsoftpdc.com

Microsoft

It Really Is Like A Utility

The New York Times
Wednesday, September 10, 2008

Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPI

Search Tech News & 8,000+ Products

Browse Products
Select a Product Category

Bits

Business » Innovation » Techno

February 15, 2008, 1:32 PM

Amazon's S3 Cloud
Start-Ups

By BRAD STONE

UPDATE: see updated Amazon st

A few days after the BlackBerry e-
giving companies another reason t
functions.

Amazon's S3 service, which offers
Web storage for hundreds of thous
companies, went down this mornin
a.m. Eastern time and is only now
back up, delivering high error rates
on Web forums.

THE WALL STREET JOURNAL.

The Online Journal
GET 2 WEEKS FREE
SUBSCRIBE NOW

As of 5:03 p.m. EDT Tuesday, June 15, 1999

News | Today's Newspaper | My Online Journal | Multimedia & Online Extras

Login | Subscribe Now – Get 2 Weeks Free

s Yahoo! Auctions Saw
e During eBay Outage

IAL NEWS ROUNDUP

panies Featured in This Article: Yahoo!, eBay

s that Yahoo! Inc.'s auction site experienced a huge spike
Bay Inc.'s day-long outage.

frustrated eBay users turned to rival services during the
ear how much long-term damage the leading Internet

lted completely for almost 22 hours, starting at 10:50 p.m.
to eBay's Web site -- one of ...

ChannelWeb

Home | Communities | Newslette

You are not logged into Channe

News | Reviews | Research | Tools | The IT Channel | Networking | Security | Storage

NEW ON CHANNELWEB

- Fast Growth Awards
- ARC Awards
- Promo Finder
- Women Of The Channel
- Apply For Tech Innovators
- CRN Fast Growth 100
- New Service: Marketing Support
- Real-Time Product Pricing and Availability
- Emerging Vendors
- VARBusiness 500

FEATURED VIDEO

ChannelWeb VIDEO

PLAY

Big Profits In a Small Package
Building relationships is

Microsoft Windows Live Services Suffer Global Outage

By Kevin McLaughlin, ChannelWeb
2:28 PM EST Tue. Feb. 26, 2008

Microsoft on Tuesday said it's looking into reports of a possible service outage affecting East Coast users of its Windows Live Hotmail, Messenger, and Skydrive services.

But the service outage could actually be much more widespread: If posters on Huliq.com forum are to be believed, the outage is not only affecting customers nationwide, but also throughout the world.

A Microsoft spokesperson said the vendor is aware that some customers may be experiencing difficulty accessing their Windows Live accounts.

"We're actively investigating the cause and are working to take the appropriate steps to remedy the situation as rapidly as possible. We sincerely apologize for any inconvenience and disruption this may be causing our customers," the spokesperson said in an email to ChannelWeb.

Microsoft has been actively pushing its 'software + services vision, in which client software and in-the-cloud services interact with each other to enhance the computing experience.

Microsoft

Climate Change and Energy

Wangari Maathai
2004 Nobel Peace Prize

Microsoft

Today's Cloud Data Centers

- Massive commodity servers
- Energy intensive infrastructure
- Cooling inefficiencies
- Environmental issues
- Expensive UPS support
- Enterprise TCP/IP networks
- Long deployment times
  - Construction and integration
- Diverse services and SLAs

- *Explosive growth*
  - *demand and expectations*

Microsoft

Consider These Services Challenges

- Environmental responsibility
  - Managing under the 100 MW envelope
  - Adaptive systems management
- Provisioning 25,000 servers
  - Hardware: at most one week after delivery
  - Software: at most a few hours
- Resilience during a blackout/disaster
  - Data center failure
  - Service rollover for 20M customers
- Programming the entire data center
  - Power, environmentals, provisioning
  - Component tracking, resilience, …

Exascale/Clouds Technical Issues

- Cooling
  - Operating points
- New pa
  - Opto
- New sto
  - Disks
- Locality
  - The
- Progran
  - Effec
- Intellige
  - Adap
- System
  - Reliability

ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems

Peter Kogge, Editor & Study Lead
Keren Bergman
Shekhar Borkar
Dan Campbell
William Carlson
William Dally
Monty Denneau
Paul Franzon
William Harrod
Kerry Hill
Jon Hiller
Sherman Karp
Stephen Keckler
Dean Klein
Robert Lucas
Mark Richards
Al Scarpelli
Steven Scott
Allan Snavely
Thomas Sterling
R. Stanley Williams
Katherine Yelick

September 28, 2008

This work was sponsored by DARPA IPTO in the ExaScale Computing Study with Dr. Wil as Program Manager; AFRL contract number **FA8650-07-C-7724.** This report is published in the interest of scientific and technical information exchange. Its publication does not constitute the Government's approval or disapproval of its ideas or findings.

NOTICE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED.

memory stacking

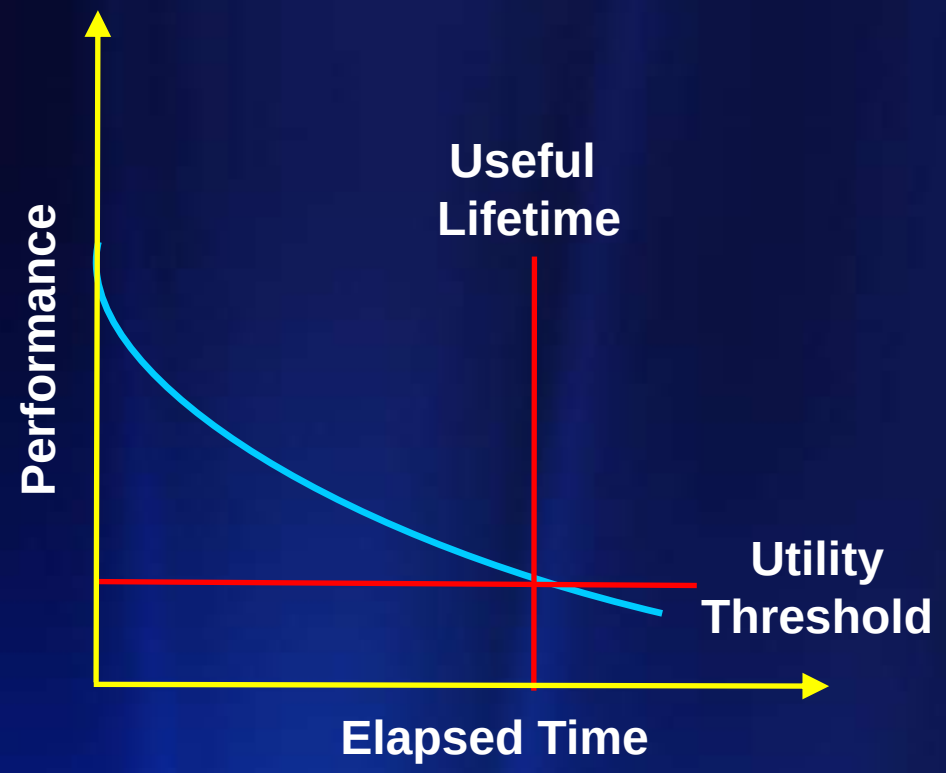
slow

scale-invariant

power down

Microsoft

Performability Thresholds



- Exascale and megadata centers
  - Design to expect failure
  - Design as energy quanta

Microsoft

Modular Data Centers

- “Come as you are” scaling
- SUN Blackbox™
- Rackable ICE Cube™
- HP and IBM

Microsoft

It's Cloud Examination Time …

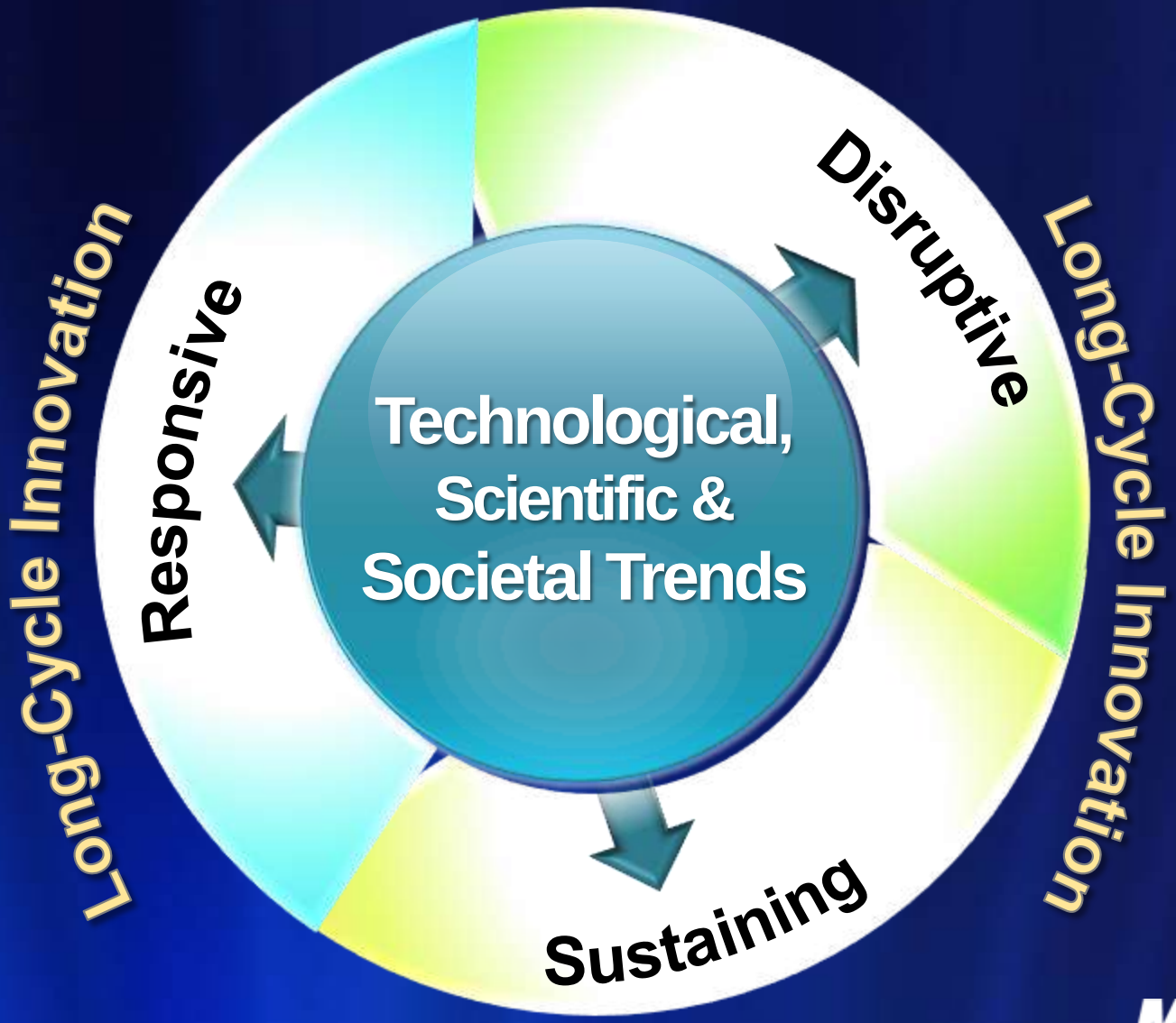
All answers must be complete, backed by rigorous mathematical, scientific and business analyses. Each incorrect response will be penalized $5B (U.S.) and loss of turn.

Q1: Compare and contrast two possible designs for an international network of data centers for web 2.0 applications that optimize energy, performance, reliability, total cost of ownership and time to market.

Q2: Repeat Q1, but for a workload of (choose one) financial services, scientific and technical applications, or rich multimedia.

Microsoft

Long-Cycle Innovation



Microsoft

Microsoft®

Your potential. Our passion.™

Microsoft®