

Energy-Efficient Data-Intensive Supercomputing

THE WORLD'S FIRST HYBRID-CORE COMPUTER.



*EnA-HPC Conference
7.-9. September 2011
Hamburg*

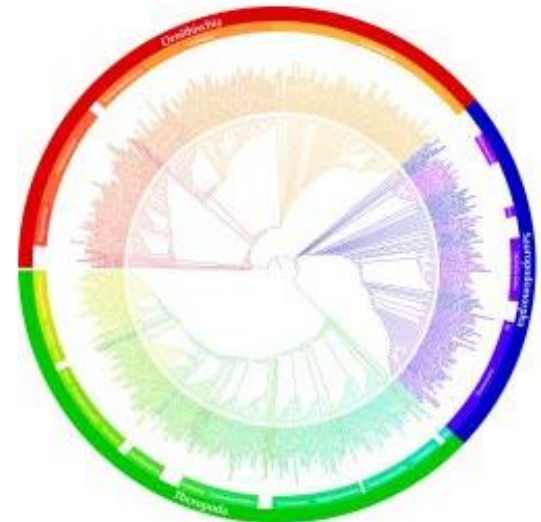
*Ernst M. Mutke
Technical Director
HMK Supercomputing GmbH*

Agenda

- **A new era of supercomputing**
- **The next computing frontier**
 - Data-intensive Supercomputing
- **Convey Architecture Overview**
- **Energy Savings Examples**

A new era of supercomputing

- **HPC is changing/growing**
 - From compute-intensive to data-intensive
- **A new class of problems**
 - Extreme data volumes
 - Complex processing
 - Highly dynamic
- **Better Energy Efficiency and Peta-Scale Computing**

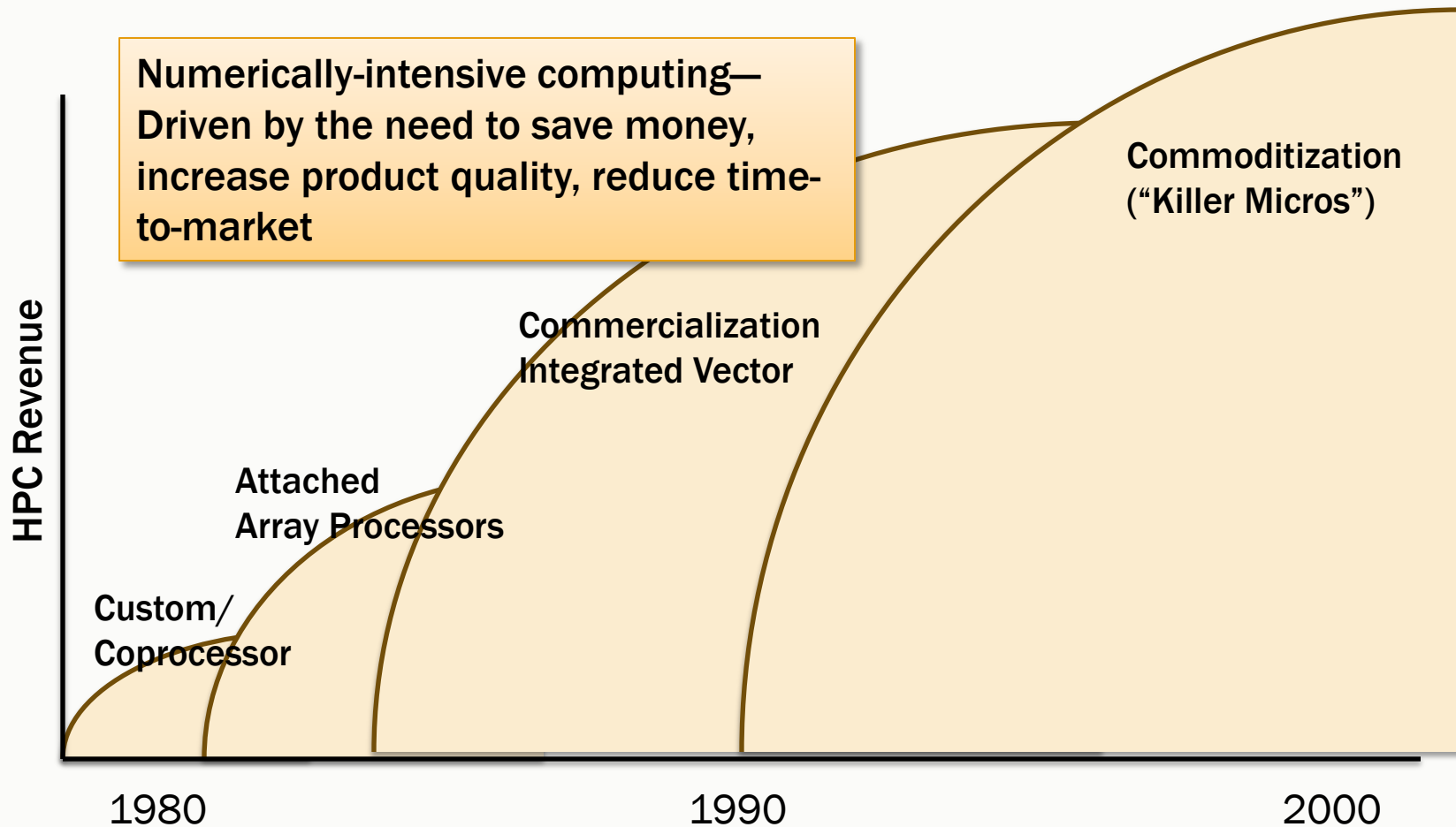


(Image: Lloyd et al/Royal Society)

“Data intensive computing demands a fundamentally different set of principles than mainstream computing.”
—National Science Foundation
Directorate for Computer and
Information Science and
Engineering

Lessons from history

The growth of numerically-intensive computing



*"The Marketplace of High Performance Computing," July 1999
Erich Strohmaier, Jack J. Dongarra, Hans W. Meuer, and Horst D. Simon

Numerically-intensive computing: Modeling real-world events

- **Used to save money, increase product quality, reduce time-to-market**
 - Computer simulation of real-world events
 - Requires FLOP/s
 - New ISA (Vector) developed
- **Required restructuring of programs**
 - New language extensions for vectorization
 - “Smart” compilers find opportunities to generate vector code
- **Ultimately supercomputers “replaced” by commodity processors**
 - Led to application-specific instructions in x86 architecture (e.g. SSE)
 - Supercomputers today are just huge clusters of x86 ISA with commodity “vector” instructions

Today: It's a data-driven world

- **Science**

- Data bases from astronomy, weather, climate, genomics, bioinformatics, natural languages, seismic modeling, ...

- **Humanities**

- Scanned books, historic documents, ...

- **Commerce**

- Corporate sales, stock market transactions, census, airline traffic, ...

- **Entertainment**

- Internet images, Hollywood movies, MP3 files, ...

- **Medicine**

- MRI & CT scans, patient records, ...

Adapted from cs.cmu.edu/~bryant

Why so much data?

- **We can produce it**
 - Automation, Internet, Sensors, Instruments
- **We can keep it**
 - Western Digital Caviar Blue 1TB - \$59.95
- **We can use it**
 - Cybersecurity
 - Medical Informatics
 - Data Enrichment
 - Social Networks
 - Symbolic Networks

“... But data-intensive applications are quickly emerging as a significant new class of HPC workloads. For this class of applications, a new kind of supercomputer, and a different way to assess them, will be required.”

—HPCwire, Nov 2010

Adapted from cs.cmu.edu/~bryant



DATA-INTENSIVE SUPERCOMPUTING

The next computing frontier: Data-Intensive Computing

- **Wal-Mart CRM**

- 267 million items/day, sold at 6,000 stores
- 4PB data warehouse
- Mine data to manage supply chain, understand market trends, formulate pricing strategies

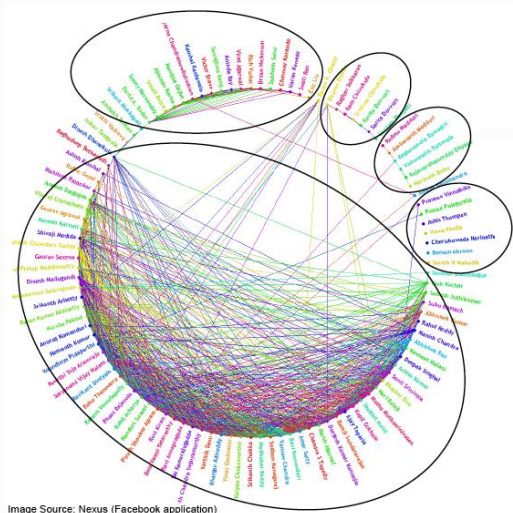
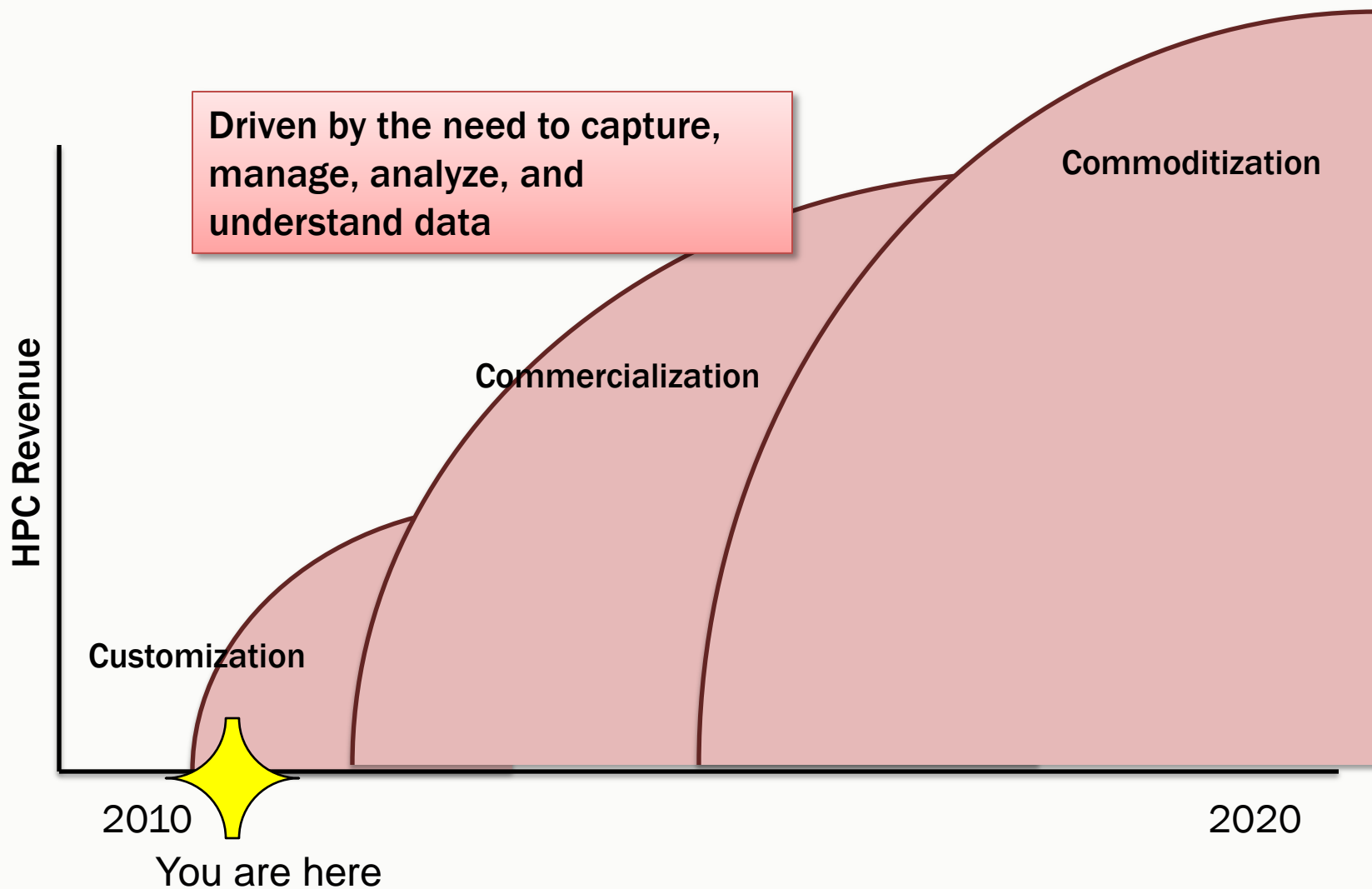


Image Source: Nexus (Facebook application)

- **Massive Social Networks**

- Detecting implicit communities, influential persons for targeted advertising

Data-intensive Computing



Data-intensive Computing

- Growing from the need to reduce computation time
- Conserve cost for energy, cooling, infrastructure, space, etc.
- Make better business decisions, reduce time-to-market
- **Requires restructuring of programs & algorithms**
 - New language extensions for MMT
 - “Smart” compilers find opportunities to generate parallel code
- **Ultimately will be “replaced” by commodity processors/systems**
 - Early data-intensive technology will be woven into mainstream processors

Architectural Characteristics

- **Reconfigurable compute elements**
 - Customizable data types
 - Application-specific logic
 - New [graph] ISA
- **Supercomputer-inspired memory subsystem**
 - Latency-tolerant
 - Large (TB's), highly-parallel memory
 - Reconfigurable architecture
 - Efficient random (cache-less) access to memory
- **Maintain x86 development ecosystem**

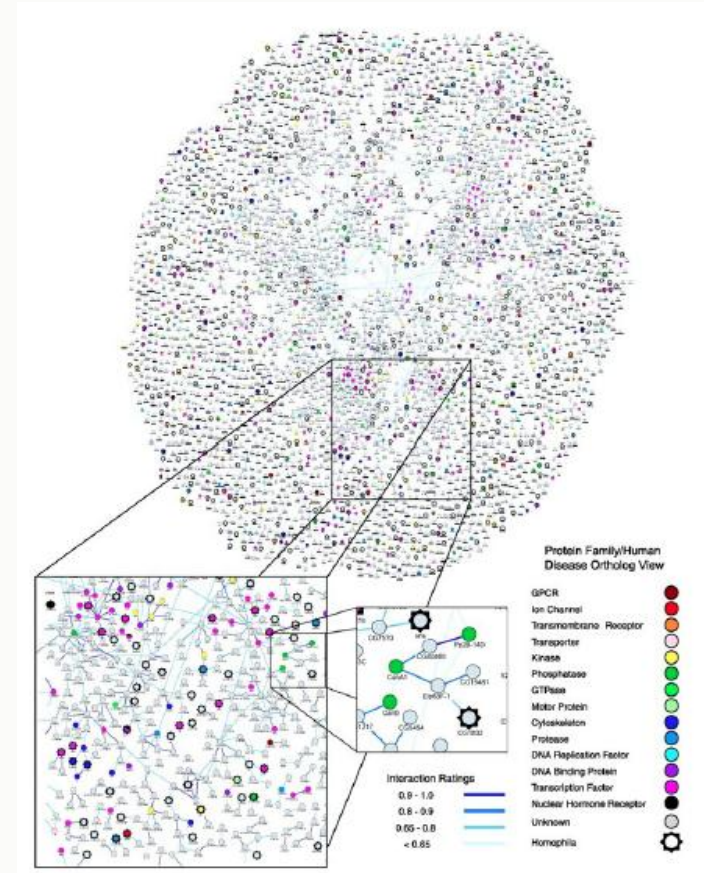
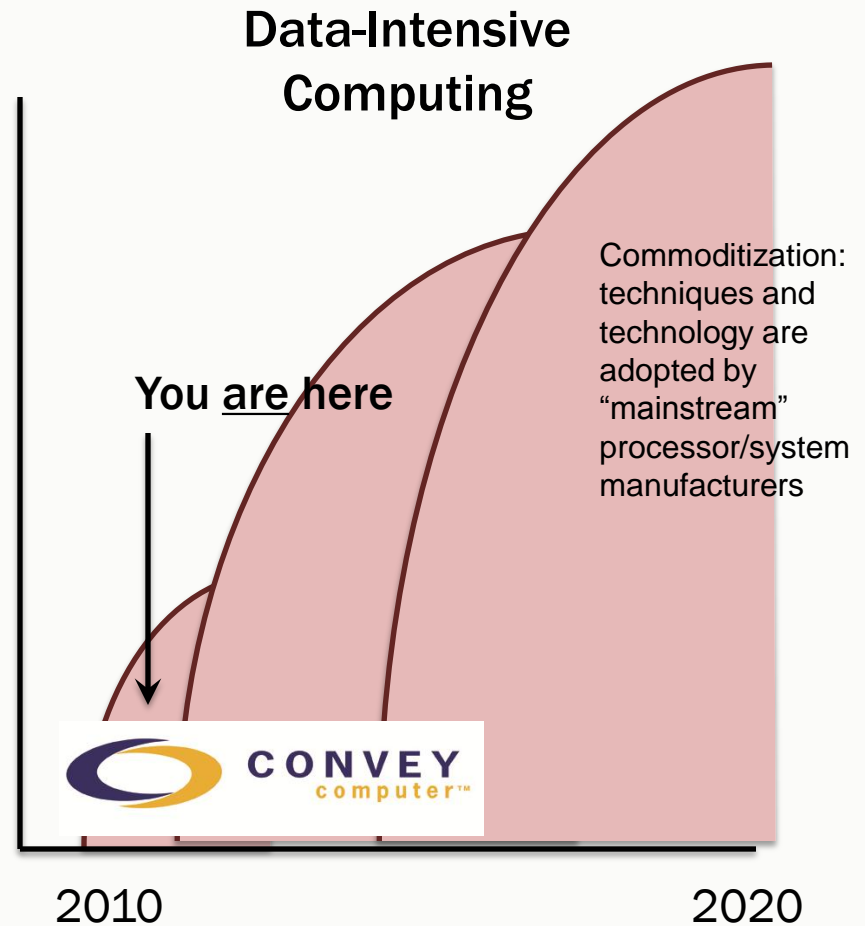
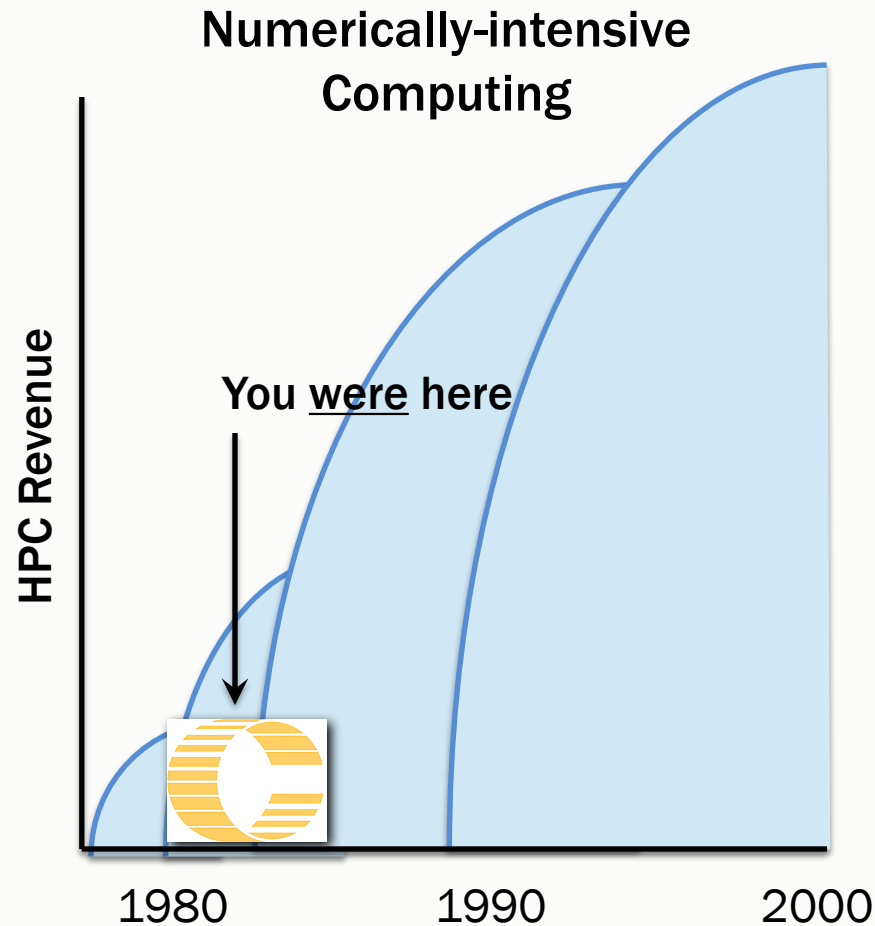


Image Source: Giotet al., "A Protein Interaction Map of *Drosophila melanogaster*", *Science* 302, 1722-1736, 2003.

Parallels



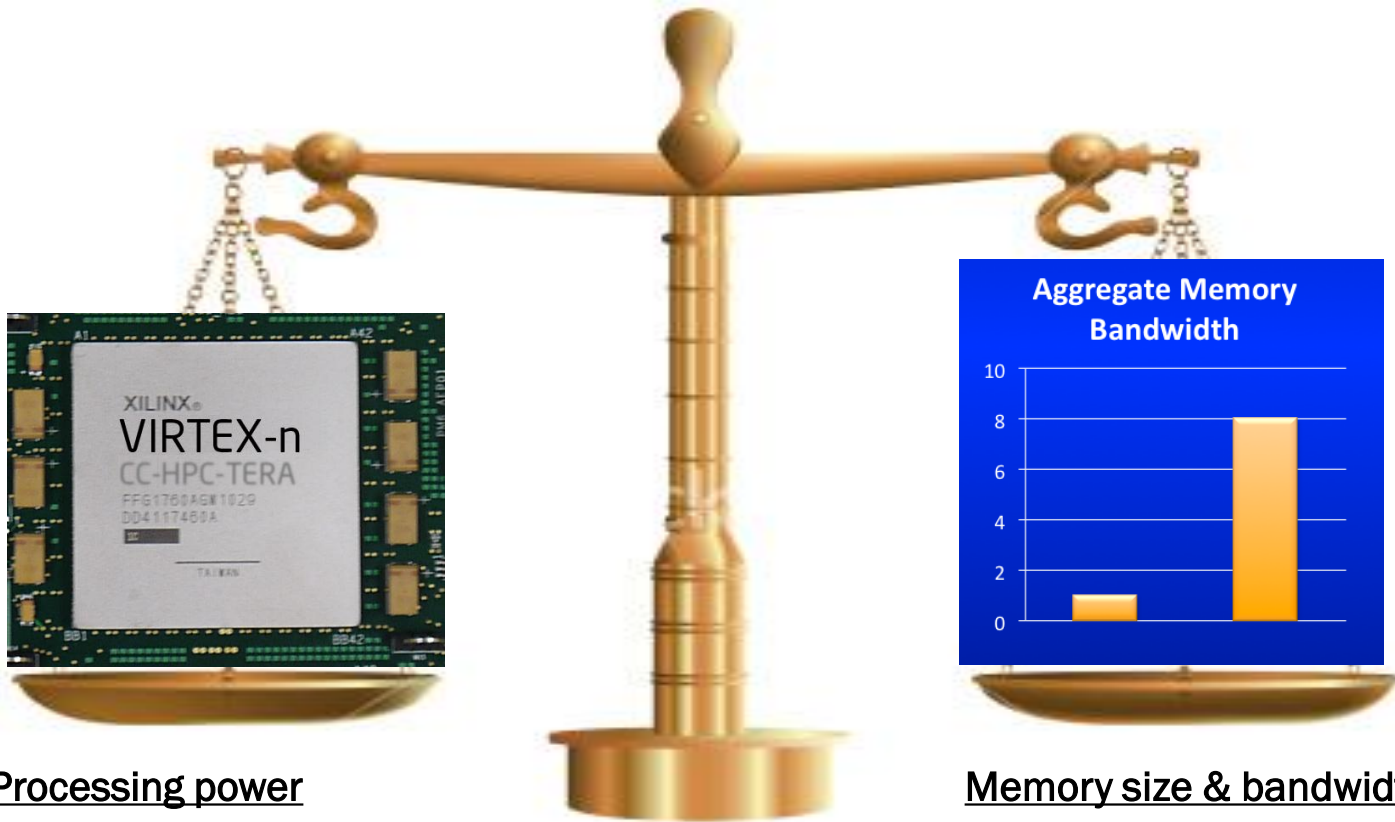


CONVEY ARCHITECTURE OVERVIEW

Design philosophies/requirements

- **Heterogeneous computing is inevitable**
 - And the simplest to program will win
 - Moore's Law is still valid, i.e. more transistors
- **Competitive/science pressures demand a different approach**
 - Must make better use of transistors
 - Support for large, randomly-accessible memory
 - Order-of-magnitude increases in performance/watt
 - Reduces OS instances, cabling, floor space, cooling requirements and power consumption
- **Convey balanced approach provides FPGA-based computing with supercomputing memory subsystems**

HPC architectures need: balanced implementations



Processing power

- Application-specific instruction sets
- Multiple techniques for parallelism (SIMD, etc.)

Memory size & bandwidth

- Highly parallel
- Atomic operations

CPU versus FPGA Comparison

- A processor executes instructions
“C” Code of 4-input logical operation

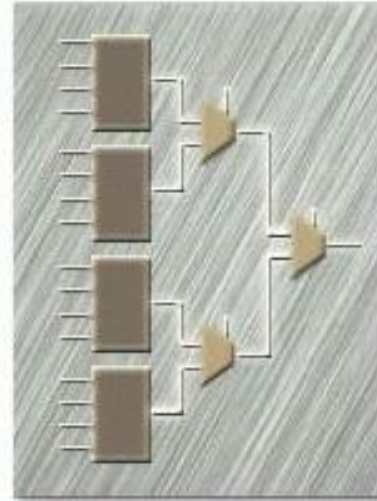
```
uint32 Log4(uint32 F, uint32 A, uint32 B,  
            uint32 C, uint32 D) {  
    uint32 R = 0;  
    for (int i = 0; i < 32; i += 1) {  
        uint32 a = (A >> i) & 1;  
        uint32 b = (B >> i) & 1;  
        uint32 c = (C >> i) & 1;  
        uint32 d = (D >> i) & 1;  
        uint32 e = (a << 3) | (b << 2)  
                | (c << 1) | d;  
        R |= ((F >> e) & 1) << i;  
    }  
    return R;  
}
```

Assembly Instructions for Log4 routine:

```
00401006 xor    edx,edx  
00401008 mov    ecx,esi  
0040100A shr    edx,cl  
0040100C and    edx,1  
0040100F lea    edi,[edx+edx]
```

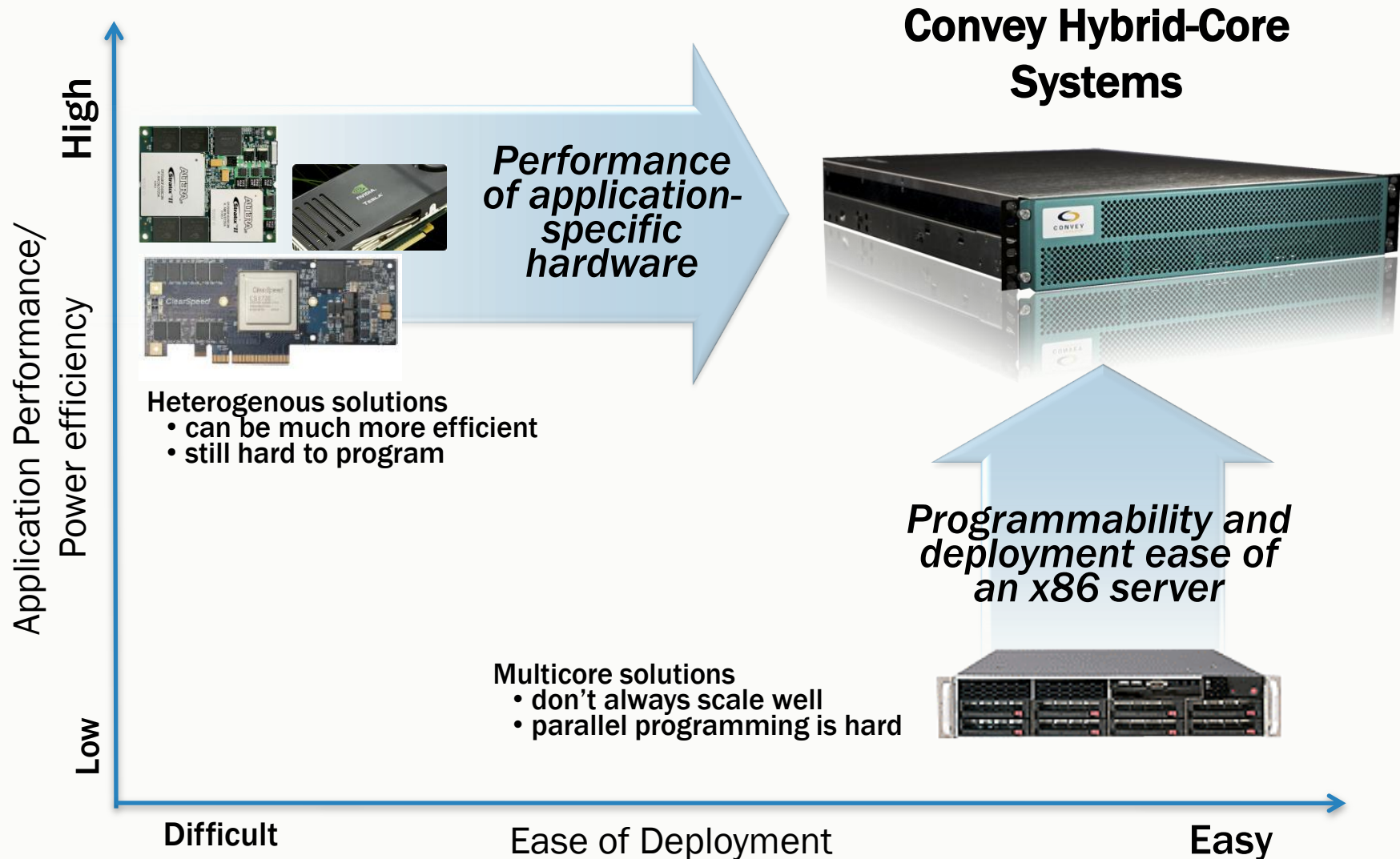
- A loop of 23 instructions are executed
32 times => 736 inst.
- 736 inst. at 3 GHz would take 245 ns
- A processor core would consume
 6.1×10^{-9} Joules (per operation)

- An FPGA uses programmable logic
FPGA Logic of 4-input logical operation

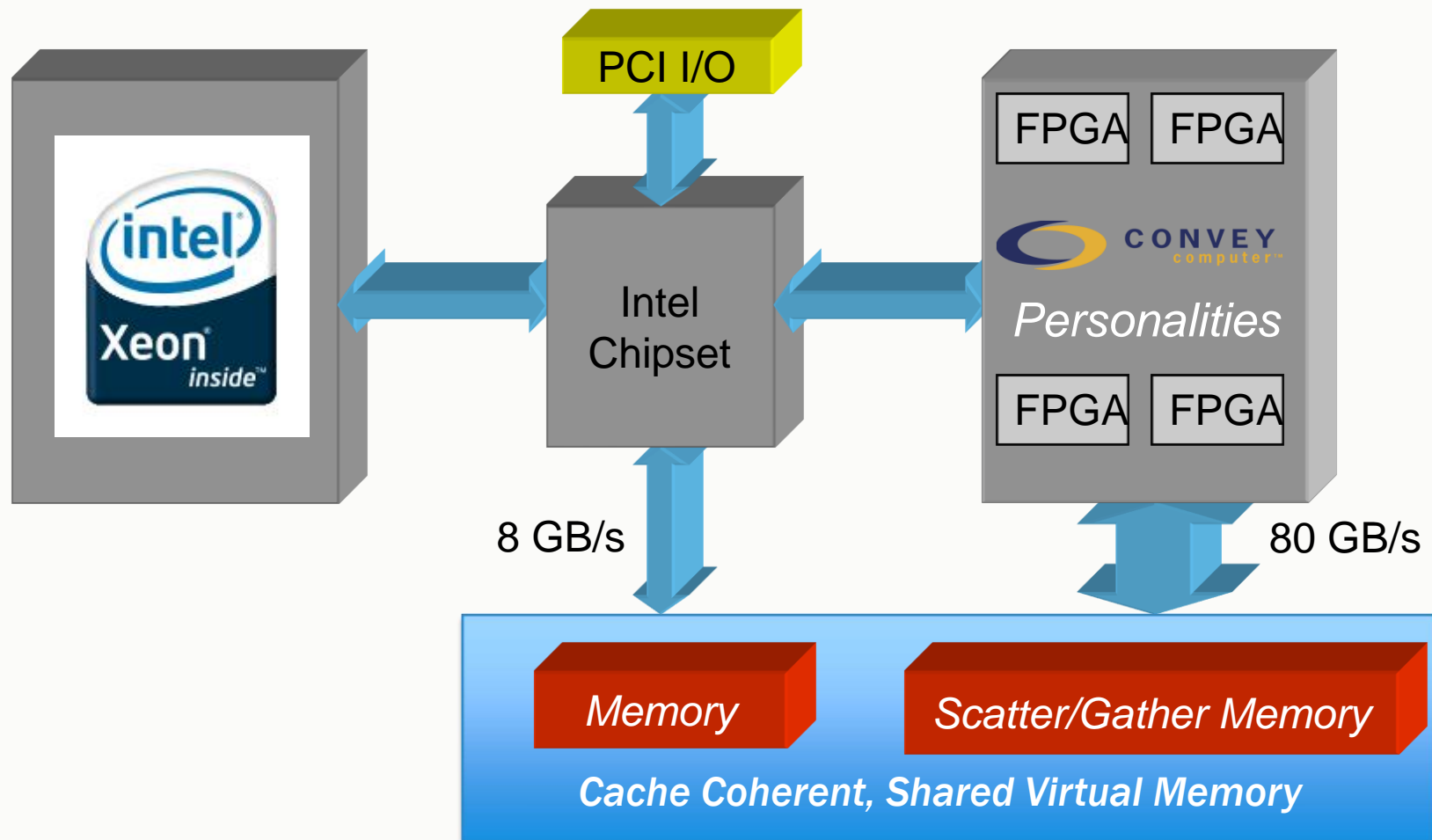


- Four logic resources per bit of result
- 32 result bits => 128 logic resources
to solve “C” routine
- The FPGA logic would take 2 ns
- An FPGA would consume 5.6×10^{-15}
Joules (per operation)

Hybrid-core Computing



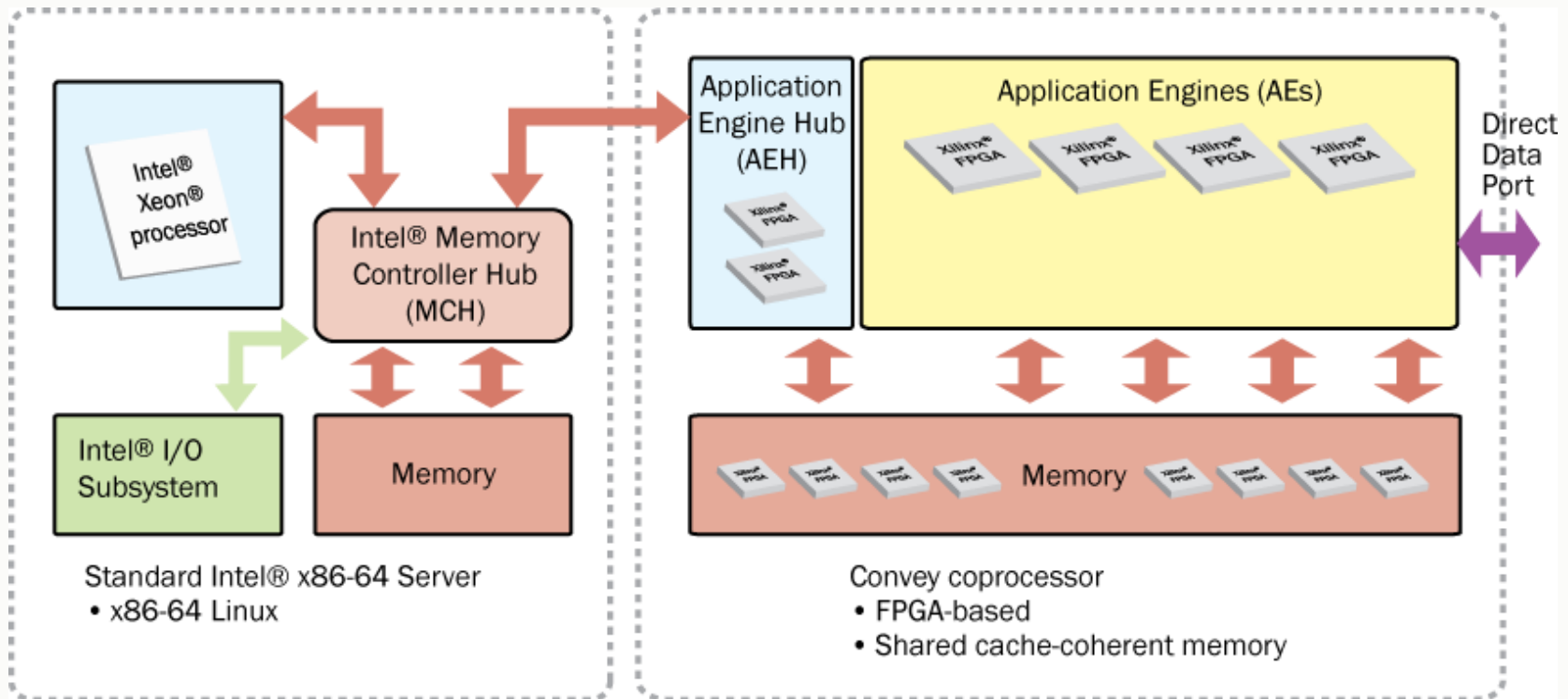
HC-1 Hardware



Convey hybrid-core architecture

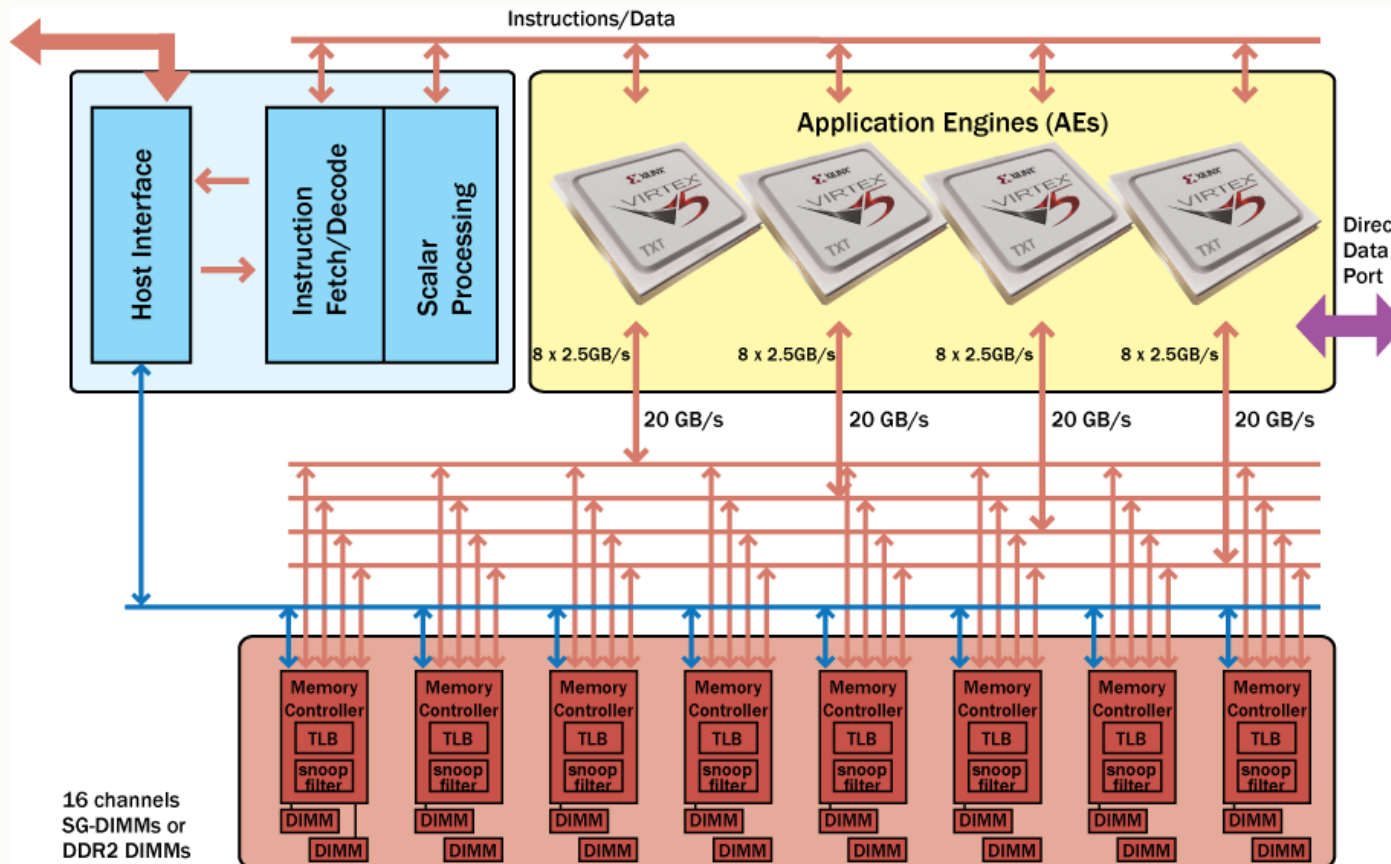
“Commodity” Intel Server

Convey FPGA-based coprocessor



Supercomputer-inspired memory subsystem

- Optimized for 64-bit accesses; 80 GB/sec peak
- Automatically maintains coherency without impacting AE performance



Random Access Memory Performance

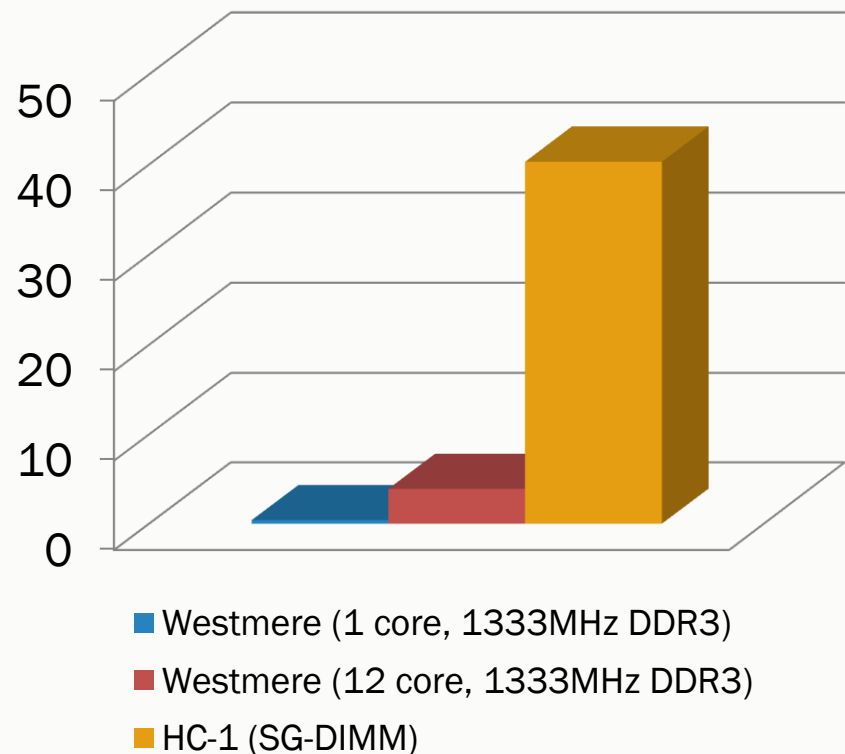
- The problem: gather elements from a large array in memory

```
for(i=0;i<nupd;i++)
```

```
Table2[i] = Table1[Index[i]];
```

- **Cache based systems are very inefficient**
 - load a whole cache line to access one element
 - random accesses to large arrays generate TLB misses
- **HC-1 coprocessor delivers a much higher percentage of peak**
 - Coprocessor memory system is designed to access 64-bit words
 - Large pages eliminate TLB misses

Gather Performance
(GB/sec)



Future Memory Requirements

- **Memory performance will continually become a larger portion of the computational bottleneck**
 - Amdahl's Law is a buzz kill when analyzing memory-bound apps... but we know this
- **Accesses that are latency sensitive [e.g., not in cache] will become much of the limiting factor**
 - As DRAM density increases, we're not doing enough creative engineering to cover the latency hot spots... more stuff through the same soda straws
- **Future algorithm and instruction set development needs to comprehend memory, computation, & programming model**
 - in order to have a reasonable chance at utilizing new core technologies
- **Flexible Memory Configuration to adopt for different memory requirements and memory access patterns**



ENERGY SAVINGS EXAMPLES

Energy Savings Examples

- **Based on performance factor**
 - calculate savings in space, energy, air conditioning costs for equivalent performance
- **Do not include savings from reducing cabling and OS instances**
- **Compares equivalent performance of Convey vs. standard x86 systems**
- **In general, compares 12core (2 x 6-core Westmere) x86 servers, but in some cases uses customer provided configurations**

Velvet/CGC (Data Intensive)

Energy comparison for equivalent performance

(1) Convey HC-1 vs Dell R910 1TB

PERF HC-1 128/64 > 5 X 4 socket 1TB Dell R910

POWER	Power Requirements[1]		
	1 racks (1 nodes) Convey	6.0	MW-h/yr
	1 racks (6 nodes) x86	73.0	MW-h/yr
	1 Year Electricity costs (@ 0.07 /kWh) [2]		
SITE	Convey	0.9	K\$/yr
	x86	10.2	K\$/yr
	1 Year Infrastructure costs[3]		
	Convey	1.9	K\$/yr
TCO	X86	18.6	K\$/yr
	3-Year TCO[4]		
	Convey	89	K\$/yr
	X86	570	K\$/yr

[1] Limit rack power to 12 kW

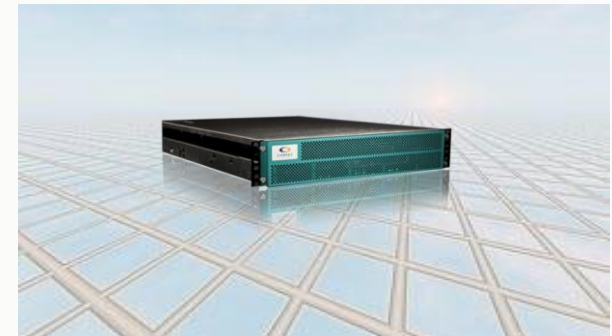
[2] Includes datacenter power/cooling costs (2x); excludes any "Green" rebates

[3] Includes prorated 10-year UPS & datacenter floorspace

[4] Includes purchase, h/w maintenance, power, infrastructure



6 x 4U 4-socket servers



1 x 2U Convey HC-1

Reduction in space	0%
Reduction in datacenter watts	91%
Reduction in 3 yr TCO	84%

Velvet/CGC (Data Intensive)

Energy comparison for equivalent performance
Convey HC-1 vs Dell R910 1TB

PERF HC-1 128/64 > 5 X 4 socket 1TB Dell R910

Power Requirements[1]			
POWER	1 racks (16 nodes) Convey	101.0	MW-h/yr
	11 racks (85 nodes) x86	1,032.0	MW-h/yr
POWER	1 Year Electricity costs (@ 0.07 /kWh) [2]		
	Convey	14.1	K\$/yr
	x86	144.4	K\$/yr
SITE	1 Year Infrastructure costs[3]		
	Convey	25.6	K\$/yr
	X86	262.1	K\$/yr
TCO	3-Year TCO[4]		
	Convey	1,386	K\$/yr
	X86	8,072	K\$/yr

[1] Limit rack power to 12 kW

[2] Includes datacenter power/cooling costs (2x); excludes any "Green" rebates

[3] Includes prorated 10-year UPS & datacenter floorspace

[4] Includes purchase, h/w maintenance, power, infrastructure



85 x 4U 4-socket servers



16 x 2U Convey HC-1

Reduction in space	91%
Reduction in datacenter watts	90%
Reduction in 3 yr TCO	83%

SWSearch (Compute Intensive)

Energy comparison for equivalent performance
Convey HC-1^{ex} vs 12-socket x86

PERF HC-1ex 32/16 \approx 10 X 12-Core 3.33 GHz x86

POWER	Power Requirements[1]		
	1 racks (8 nodes) Convey	50.0	MW-h/yr
	3 racks (77 nodes) x86	233.0	MW-h/yr
	1 Year Electricity costs (@ 0.07 /kWh) [2]		
POWER	Convey	7.1	K\$/yr
	x86	32.6	K\$/yr
SITE	1 Year Infrastructure costs[3]		
	Convey	12.9	K\$/yr
	X86	59.3	K\$/yr
TCO	3-Year TCO[4]		
	Convey	578	K\$/yr
	X86	1,184	K\$/yr

[1] Limit rack power to 12 kW

[2] Includes datacenter power/cooling costs (2x); excludes any "Green" rebates

[3] Includes prorated 10-year UPS & datacenter floorspace

[4] Includes purchase, h/w maintenance, power, infrastructure



77 x 1U 12-core servers



16 x 3U Convey HC-1^{ex}

Reduction in space	67%
Reduction in datacenter watts	78%
Reduction in 3 yr TCO	51%

PCAP (Data & Compute Intensive)

Energy comparison for equivalent performance
Convey HC-1 vs 2-socket 8-core x86

PERF

HC-1 32/16 > 111 X 2 socket 8-core x86

POWER

Power Requirements[1]

1 racks (16 nodes) Convey 101.0 W-h/yr

53 racks (1775 nodes) x86 5,364.0 W-h/yr

1 Year Electricity costs (@ 0.05 /kWh) [2]

Convey 10.1 K\$/yr

x86 536.4 K\$/yr

SITE

1 Year Infrastructure costs[3]

Convey 25.6 K\$/yr

X86 1,361.7 K\$/yr

TCO

3-Year TCO[4]

Convey 996 K\$/yr

X86 19,086 K\$/yr



1,775 x 1U 8-core servers



16 x 2U Convey HC-1

Reduction in space	98%
Reduction in datacenter watts	98%
Reduction in 3 yr TCO	95%

[1] Limit rack power to 12 kW

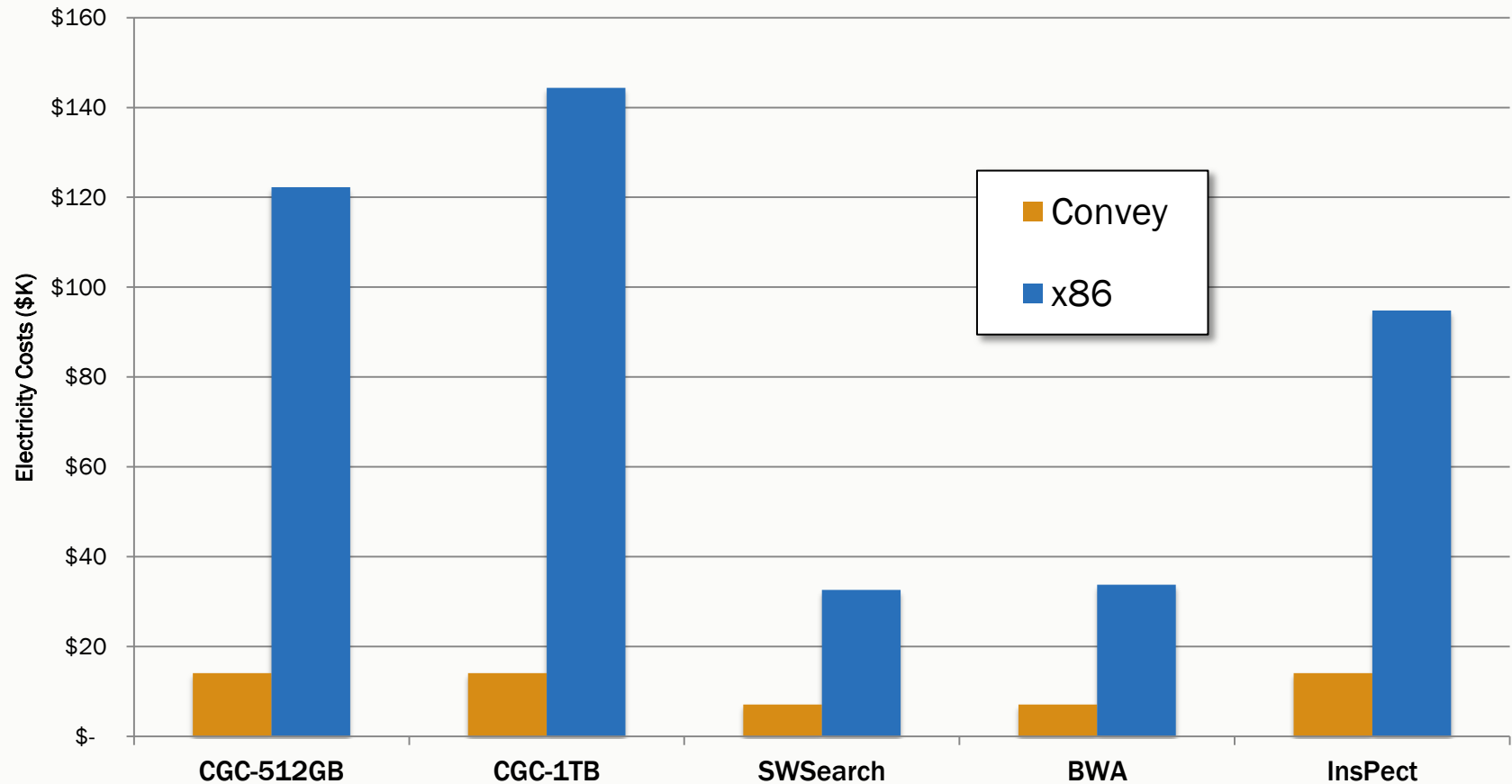
[2] Includes datacenter power/cooling costs (2x); excludes any "Green" rebates

[3] Includes prorated 10-year UPS & datacenter floorspace

[4] Includes purchase, h/w maintenance, power, infrastructure












Electricity Cost Comparison

1 Year Electricity costs



*Includes datacenter power/cooling costs @ \$.07/KWh; excludes any "Green" rebates

Graph500: Performance Rank (Problem Scale 31 and lower)

Rank	System	Site	Scale	MTEPS	Perf/ W
13	SGI Altix ICE 8400EX, 256 nodes / 1024 cores	SGI	31	14,085	 363
14	NNSA/SC Blue Gene/Q Prototype II (512 nodes)	IBM Research, T.J. Watson	31	11,323	 362
15	DAS-4/VU (SuperMicro, 64 nodes / 512 cores)	VU University	31	4,642	 91
18	SuperDragon-1 (Sugon, 32 nodes / 384 cores)	Inst of Computing Tech, Beijing	30	1,454	-
21	cougarxmt (Cray XMT, 128 nodes)	PNL	29	1,223	 12
22	graphstorm (Cray XMT, 128 nodes)	SNL	29	1,171	12
-	Vortex (Convey HC-1ex, 1 node / 4 cores, 4 FPGAs)	Convey Computer Corporation	27	1,122	 1,496
19	Jaguar (Cray XT5-HE, 18,688 nodes / 224,256 cores)	ORNL	30	1,011	0
16	Matterhorn (Cray XMT2, 64 nodes)	CSCS	31	885	 18
23	Matterhorn (Cray XMT2, 64 nodes)	CSCS	29	879	 18
28	Minerva (IBM iDataPlex, 258 nodes / 3096 cores)	University of Warwick	26	839	-
26	Vortex (Convey HC-1ex, 1 node / 4 cores, 4 FPGAs)	Convey Computer Corporation	27	773	 1,031
27	Westmere E7-4870 2.4GHz, 1 node / 40 cores	Intel Research	27	705	 320
24	Erdos (Cray XMT, 64 nodes)	ORNL	29	702	 14
20	Knot (HP MPI cluster, 8 processors / 64 cores)	UCSB	30	177	9
17	Kraken (Appro, 1 node / 32 cores)	LLNL	31	105	 75
29	Neumann (HPC Systems, 32 cores)	UCSB	26	40	6
25	Gordon (Appro, 7 nodes / 84 cores)	SDSC	29	30	3

Observations & Conclusions

- **HPC is changing/growing**
 - Data-intensive applications are a must for industry
 - Heterogeneous (hybrid) systems are inevitable
- **It looks a lot like 1980**
 - New architectures to address the challenges of new computing requirements
 - Early adopters establish standards & technology
- **Current commodity architectures are not suitable for data intensive jobs**
 - Memory subsystems, access pattern and data location
- **Need better scalability and cost savings for future data intensive challenges**
 - Energy, Cooling, Space, Infrastructure



THANK YOU!

Questions??