



PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE

Emerging and future technologies

Herbert Huber, DEISA PRACE Symposium 2010



Outline

- Emerging Technologies (PRACE WP8) prototypes
- Future Technologies (The Path to Exascale)
 - System Architecture
 - Supercomputing Site Infrastructures
 - Memory
 - Processing Units
 - Interconnect
- Summary

Emerging Technologies: WP8 Prototypes I



CEA
“GPU/CAPS”

1U Tesla Server T1070 (CUDA, CAPS, DDT) Intel Harpertown nodes

Take more easily advantage of accelerators. Compare HMPP with other approaches to program accelerators.

CINECA I/O Subsystem (SSD, Lustre, pNFS)

Assess the applicability of new file system and storage technologies.

CINES-LRZ
“LRB/CS”

Hybrid SGI ICE/UV/Nehalem-EP & Nehalem-EX/ClearSpeed/Larrabee

Evaluate of a hybrid system architecture containing thin nodes, fat nodes and compute accelerators with a shared file system.

CSCS
“UPC/CAF”

Prototype PGAS language compilers (CAF + UPC for Cray XT systems)

Understand usability and programmability of a PGAS language.



EPCC
“FPGA”

Maxwell – FPGA prototype (VHDL support & consultancy + software licenses (e.g., Mitrion-C)

Assess the potential of high-level languages for using FPGAs in HPC.
Compare energy efficiency with other solutions.

Emerging Technologies: WP8 Prototypes II

FZJ “Cell & FPGA interconnect”

eQPACE (PowerXCell
8i cluster with special
network processor)

Gain deep expertise in communication
network issues.
Extend the application domain of the
QPACE system.

LRZ “RapidMind”

RapidMind Multicore Development
Platform (automatic code generation
for x86, GPUs and Cell)

Assess the potential of data stream languages.
Compare RapidMind with other approaches for
programming accelerators or multicore systems

NCF “ClearSpeed”

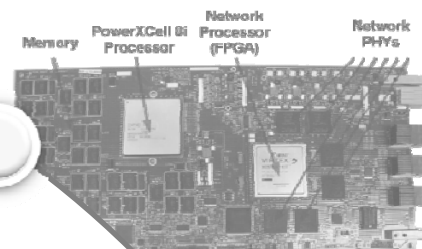
ClearSpeed
CATS 700 units

Evaluate ClearSpeed accelerator hardware
for large-scale applications.

SNIC- KTH

Air cooled blade system from
Supermicro with AMD Istanbul
processors & QDR IB
(*subject to EC approval*)

Evaluate and optimize energy
efficiency and
packing density of commodity
hardware.



PARTNERSHIP
FOR ADVANCED COMPUTING
IN EUROPE



Future Technologies: The Path to Exascale

Future System Architecture Projections/Targets

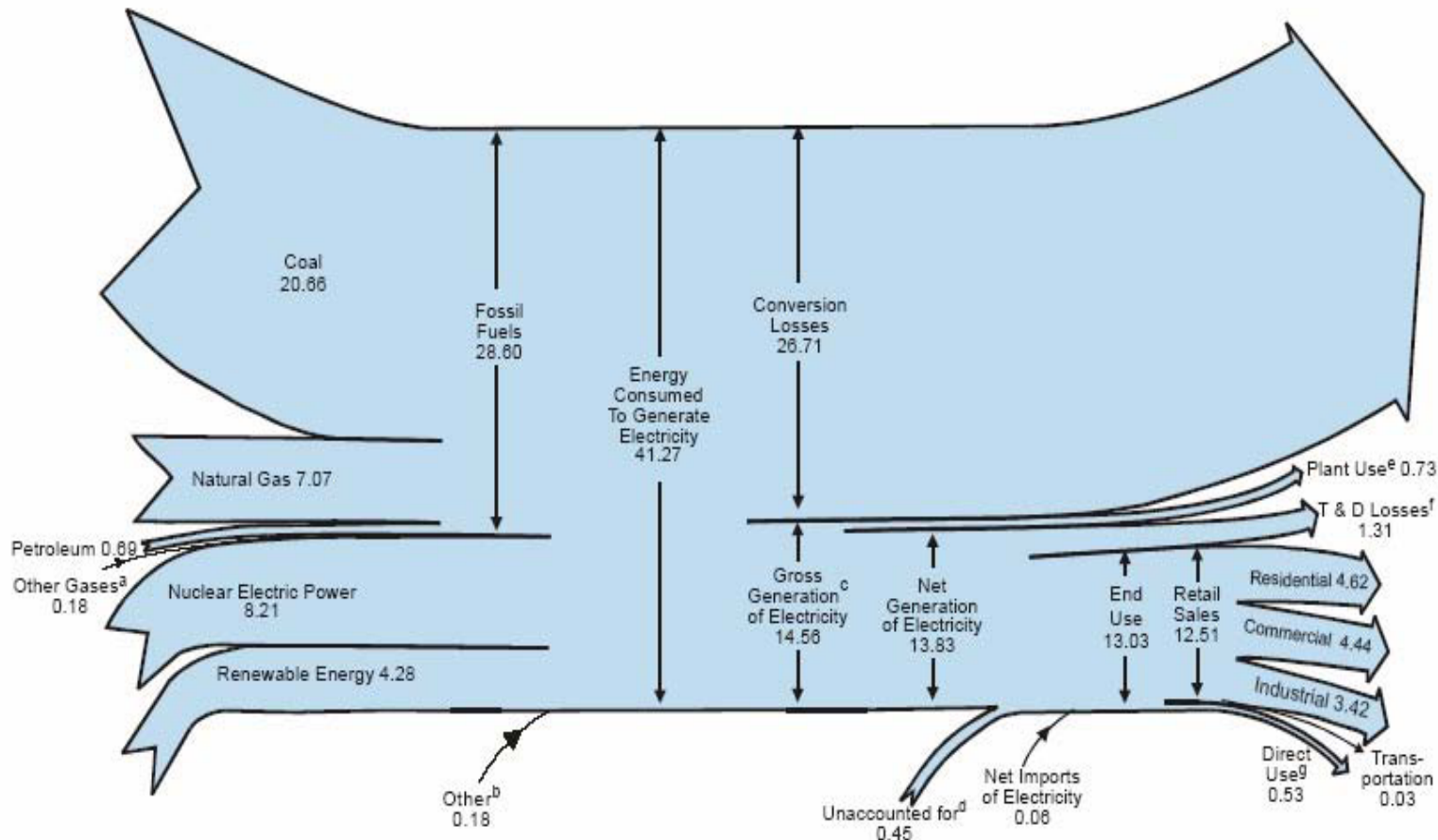
Systems	2010	2012	2015	2018	Difference Today & 2018
System Peak [PF]	2	25	200	1000	O(1000)
Power [MW]	6	6-20	15-50	20-80	O(10)
System Memory [PB]	0.3	0.3-0.5	5	32-64	O(100)
GB RAM/Core	0.5-4	0.5-2	0.2-1	0.1-0.5	-O(10)
Node Performance [GF]	125	160-1000	500-7000	1000-10000	O(10)-O(1000)
Cores/Node	12	16-32	100-1000	1000-10000	O(100)-O(1000)
Node memory BW [GB/s]	40	70	100-1000	400-4000	O(100)
Number of nodes	~20000	10.000- 100.000	5000- 50.000	100.000- 1.000.000	O(10)-O(100)
Total concurrency	~240000	O(10 ⁶)	O(10 ⁷)	O(10 ⁹)	O(10.000)
MTTI	days	days	O(1 day)	O(1 day)	-O(10)

Source: Rick Stevens and Andy White, IESP Meeting, Oxford 2010

The Path to Exascale

- System power is a first class constraint for the multi Peta- to Exascale performance regime → Reducing power is fundamentally
 - Memory (2x-5x) ↓
 - New memory interfaces (3D stacking)
 - Replace DRAM with zero power non-volatile memory
 - Processor (10x-20x) ↓
 - Reduction of data movement (> 20x)
 - Domain/Core power gating and aggressive voltage scaling
 - Interconnect (2x-5x) ↓
 - Replace copper with integrated optics
 - Supercomputing site energy efficiencies (2x) ↑
 - Free cooling of compute nodes (new system packaging and cooling technologies)
 - Onsite tri-generation of heat & district heating and cooling

Conventional Energy Efficiency



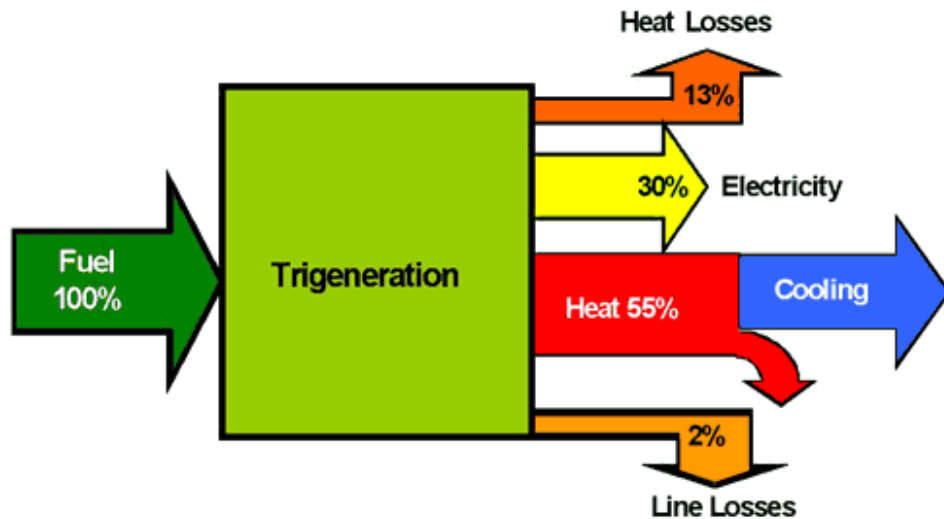
Power plant efficiency:
~35%

Transportation losses:
6%

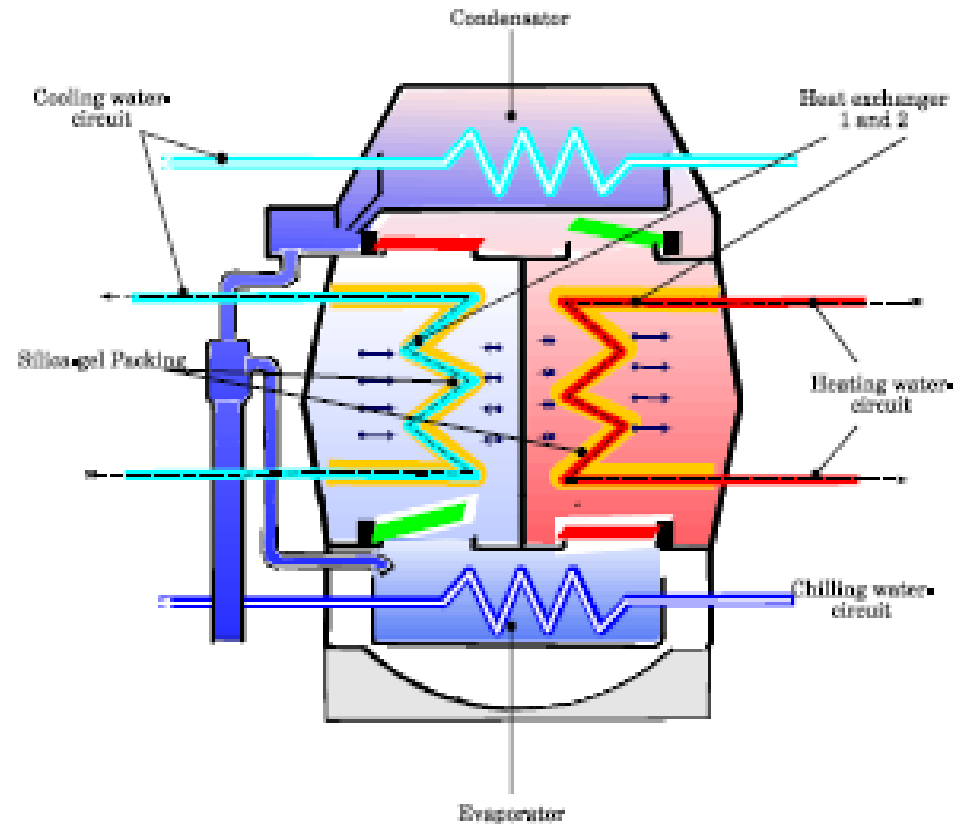
→ 70% of fuel energy
is lost!

**Energy efficiency of
whole supply chain
must be largely
enhanced**

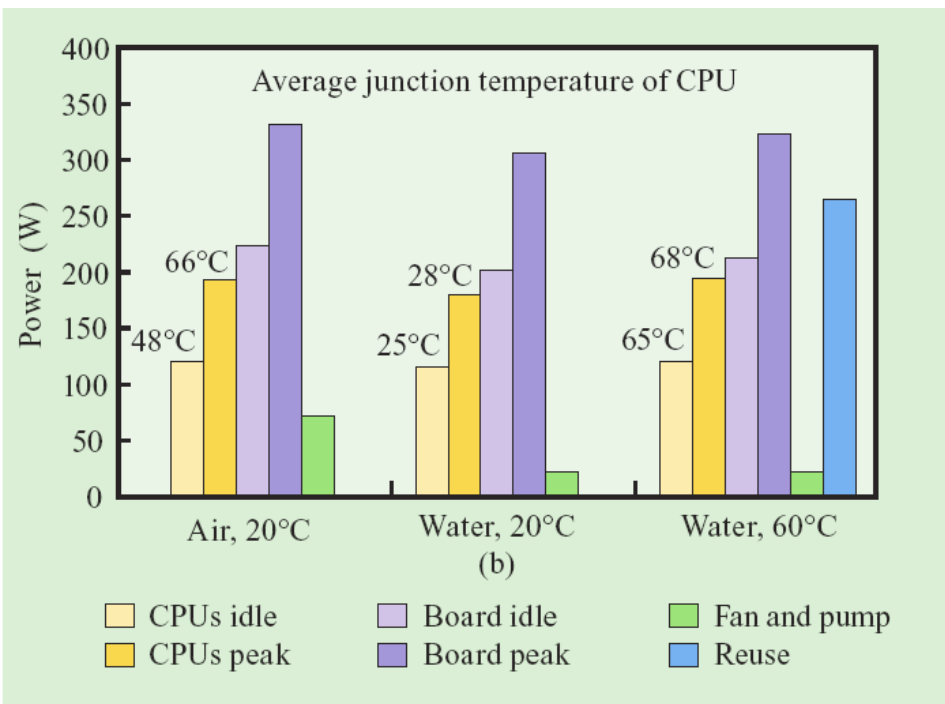
Means to enhance energy efficiency of supply chain: Tri-generation



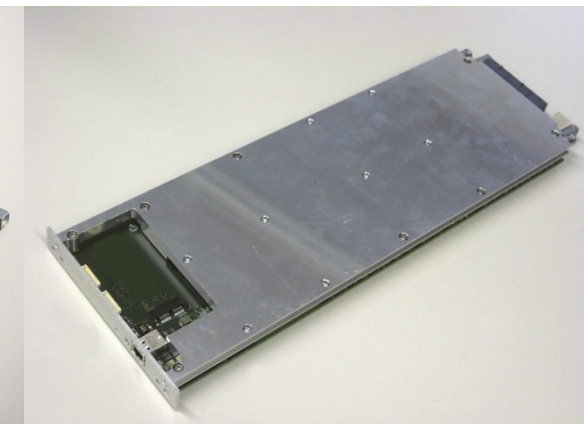
Efficiency: 85%



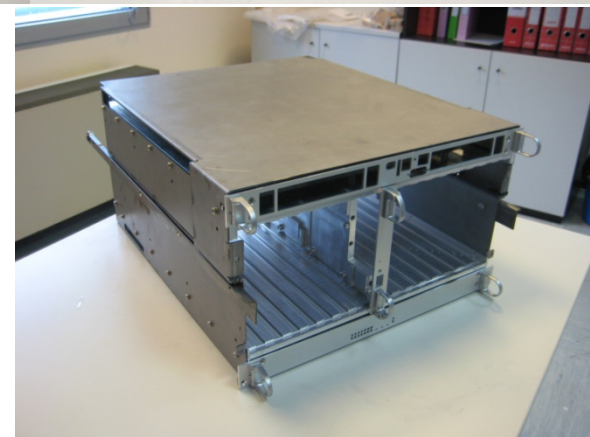
Putting it all together: Tri-generation (with district heating & cooling) & direct cooling of chips with warm water



Source: B. Michel, IBM Research Zürich, 2009

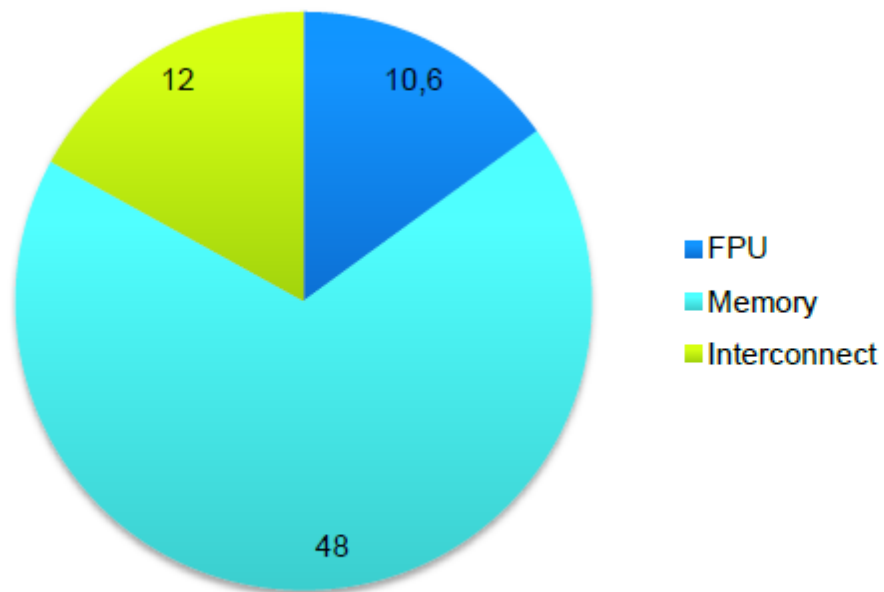


Source: G. Tecchioli, Eurotech, 2009



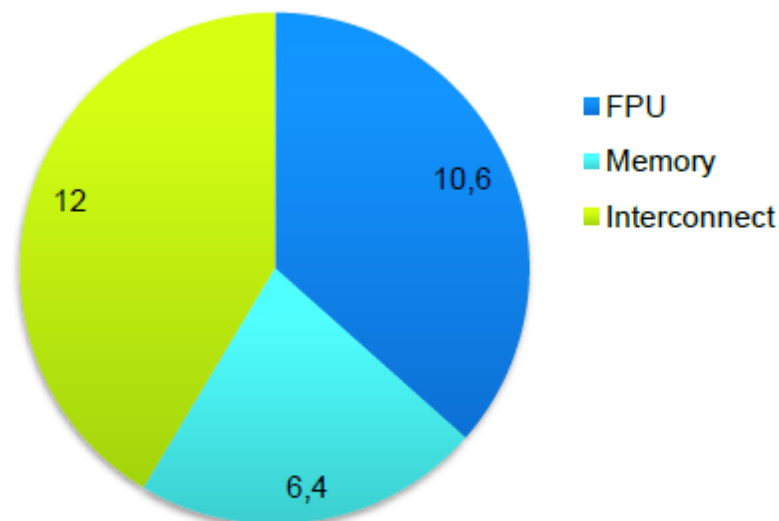
Memory and Power Consumption

Standard Memory
Technology Roadmap



70 Megawatts total

Investment in Advanced
Memory Technology

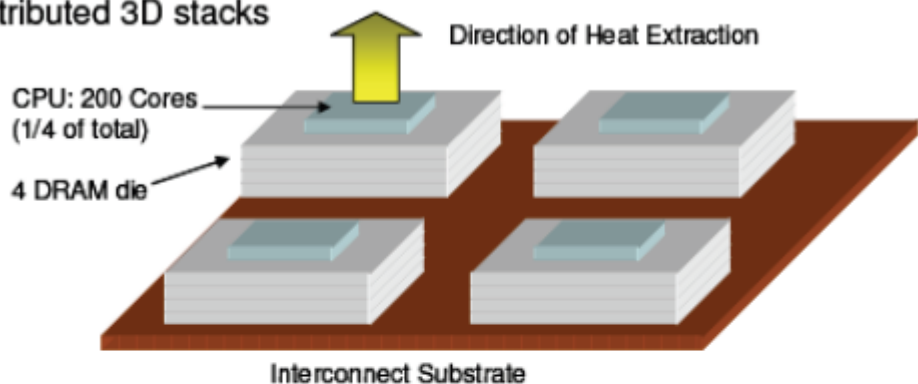


20 Megawatts total

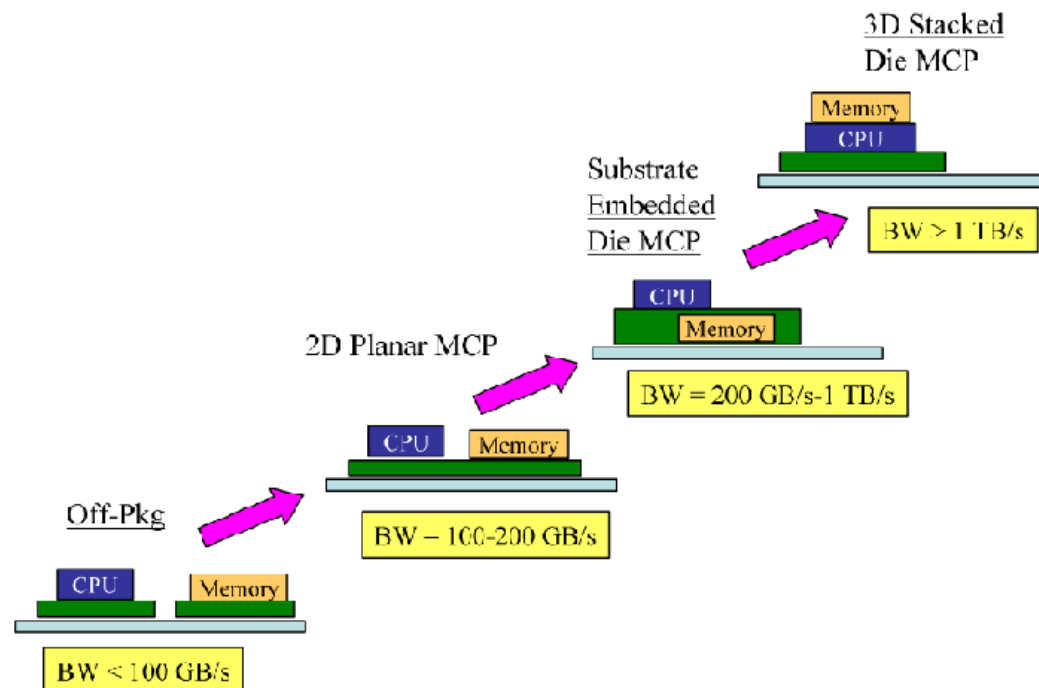
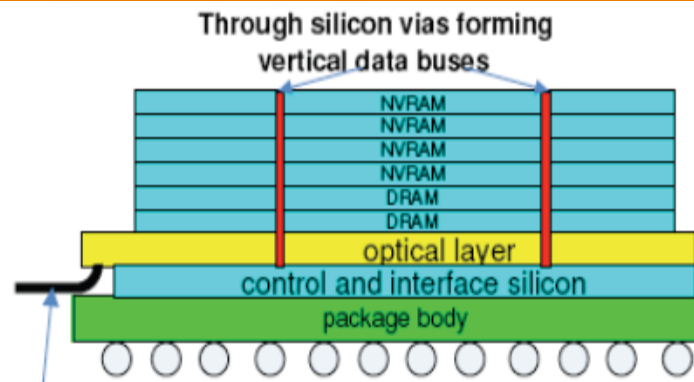
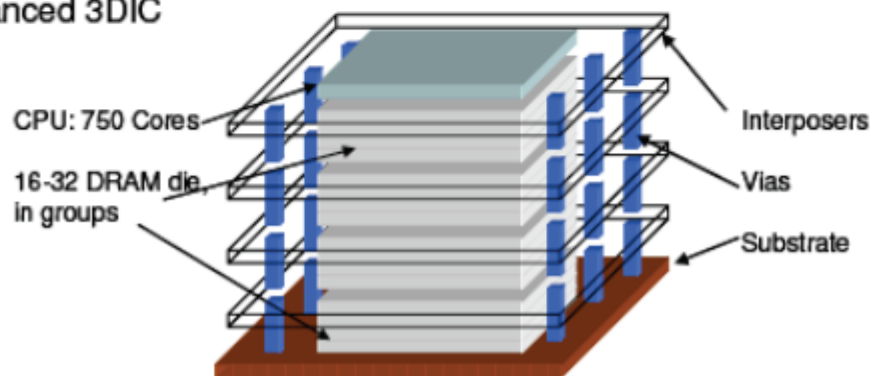
Memory: 3D Packaging Options

Approach

Distributed 3D stacks

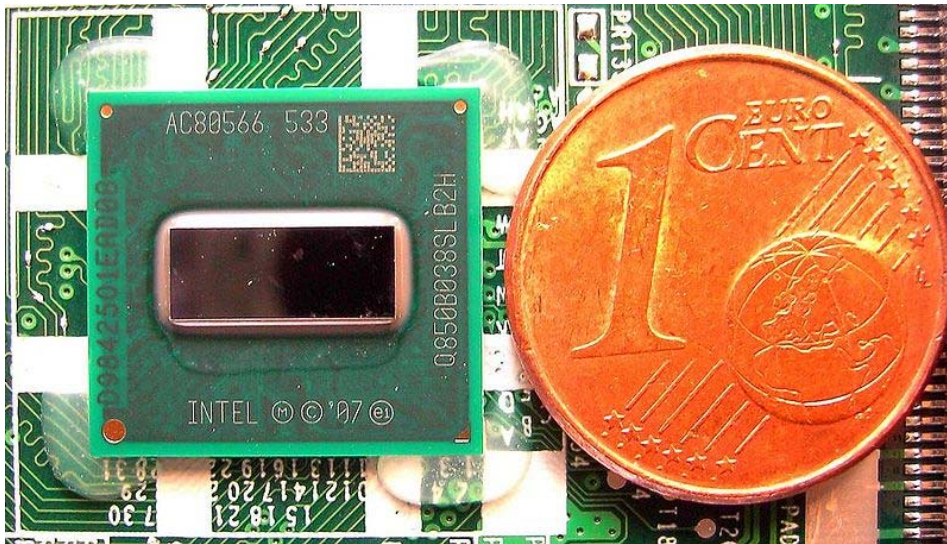


Advanced 3DIC



Processing Units: Processor Design for Low Power

Lessons to learn from embedded processor design



- Intel Atom Z500: 0.65 W
- Tensilica DP: 0.09 W

- Cubic power improvement with lower clock rate due to V^2F
↓
- Slower clock rates enable use of simpler cores
↓
- Simpler cores use less area and reduce cost
↓
- Tailor design to application

Interconnect: Server Interconnect Hierarchy

WAN, MAN
Internet, GRID



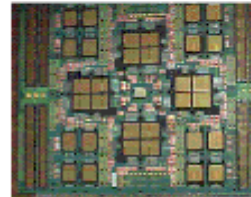
Lab- & Campus-Level
LAN, SAN



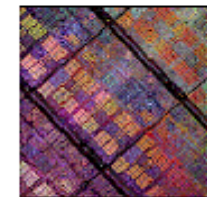
System-Level
Intra-Rack & Rack-to-rack



Board-level
Module-to-Module & Chip-to-Chip



Chip-Level
On-chip & Chip-to-Chip

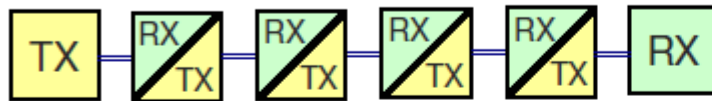


	Outside of the Box		Rack Level	In the Box	
Distances	10s to 100s km	10m to 2km	<2m (intra) to <100m (R2R)	1cm to 100cm	>3 cm
# of Lines	Singles	Tens	100s to 1000s	1000s	10000s
Standards-based or design-specific?	Internet protocol, SONET, ATM	LAN/SAN Standards (Ethernet, Fibre Channel, InfiniBand)	Moving to standards-based (InfiniBand, Ethernet)	Design-specific, some standards (PCIe, HT, QPI)	Design-specific
Optics or Copper?	Optics ubiquitous since 70s & 80s	Optics common since 90s (1/10GE, FC, IB)	Rack-to-Rack optics: now Intra-Rack optics: '10-11	Optics cost effective vs. copper in ~2012-2013	Optics probably ~2015-2017

Interconnect: Moving to All-Optical High-Speed I/O

Replace off-chip electrical drivers and on-chip electrical network with a single optical network performing serving all on/off-chip data I/O

Electronic Network



- **On chip**
 - buffer, receive and re-transmit every single bit at every switch
- **Off-chip**
 - very power hungry due to 50 Ω line drivers with heavy TX + RX equalization
 - bandwidth limited by pin count
- **Power is bandwidth x length dependent**

Optical Network



- **On and Off-chip I/O**
 - Modulate/receive ultra-high bandwidth data stream once – no re-transmit
 - Off-chip and on-chip power and bandwidth are equivalent
 - Broadband switch fabric is nearly free in power dissipation \rightarrow highly scalable
- **Power independent of bit rate and length**

Power efficient computing \rightarrow More FLOPs/W for Optical Network

Summary

- **Future applications must scale to $O(10^9)$ of processing cores**
- Power is a huge problem in HPC
- Highly likely all future supercomputers will be direct liquid cooled
- Huge technology investments (memory, processor and interconnect) are needed to reach the ExaFlop/s performance regime
 - Be prepared for low Byte/Flop ratios
 - Memory power and I/O power are biggest problems for Exascale (memory and I/O can consume up to 80 MW)



Thanks for Your attendance!

Questions?