Implications of Memory-Centric Computing Architectures for Future NoCs

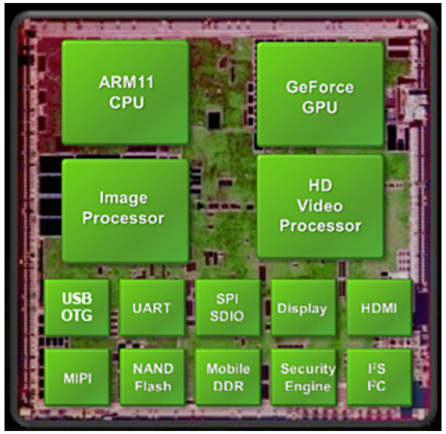
Sudhakar Yalamanchili

Computer Architecture and Systems Laboratory
Center for Experimental Research in Computer Systems
School of Electrical and Computer Engineering
Georgia Institute of Technology

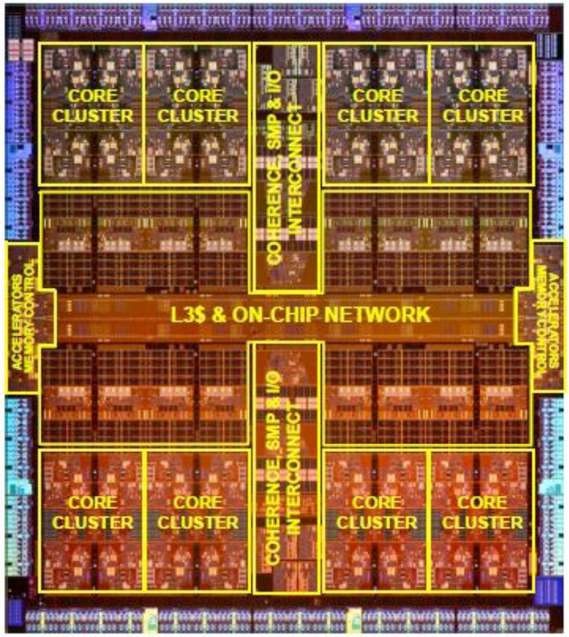
Sponsors:

Role of NoCs

Nvidia Tegra4

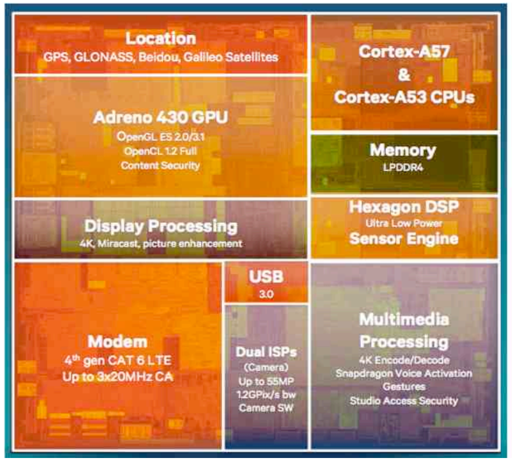


IBM Power8


CORE CLUSTER
CORE CLUSTER
COHERENCE, SMP & I/O INTERCONNECT
CORE CLUSTER
CORE CLUSTER
ACCELERATORS MEMORY CONTROL
L3$ & ON-CHIP NETWORK
ACCELERATORS MEMORY CONTROL
CORE CLUSTER
CORE CLUSTER
COHERENCE, SMP & I/O INTERCONNECT
CORE CLUSTER
CORE CLUSTER

www.theregister.co.uk

Qualcomm Snapdragon



www.themobileindian.com

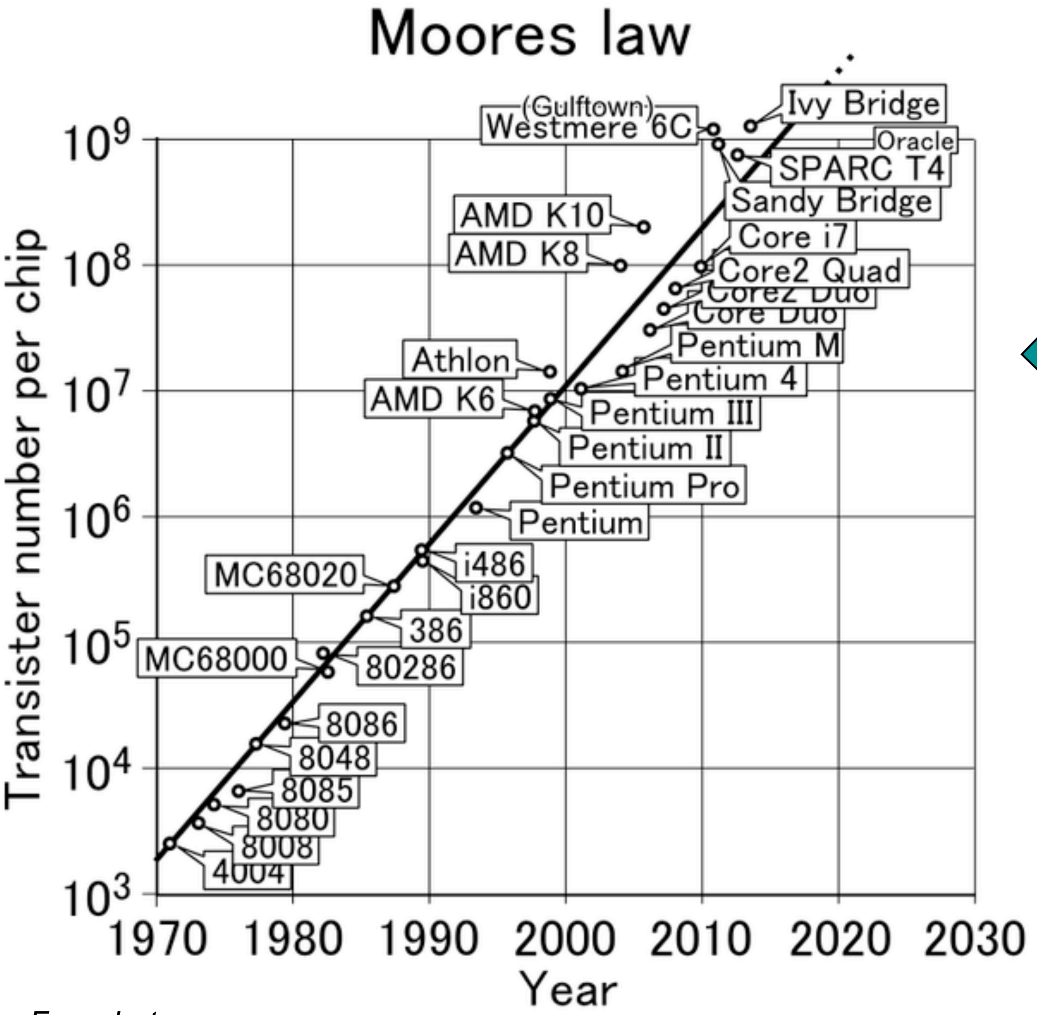
The System Defines the NoC Requirements

Overview

- Impact of Technology and Applications
- Transition to Memory Centric Compute: Inside the Package
- Transition to Memory Centric Compute: Inside the Stack
- Concluding Remarks

How are Technology and Applications Reshaping Systems?

Moore’s Law and the End of Dennard Scaling



From betanews.com

<u>Goal</u>: Sustain Performance Scaling

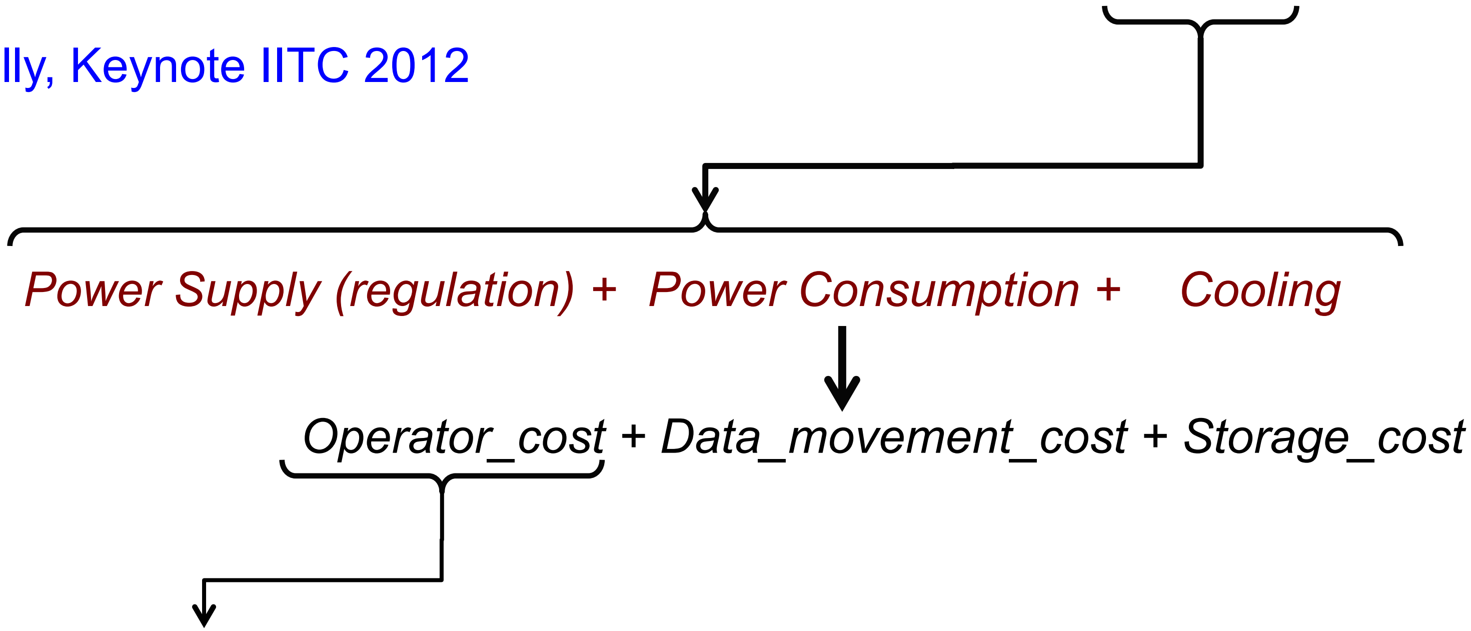
- Performance scaled with number of transistors*

*R. Dennard, et al., “Design of ion-implanted MOSFETs with very small physical dimensions,” *IEEE Journal of Solid State Circuits*, vol. SC-9, no. 5, pp. 256-268, Oct. 1974.

Power and Performance

$$Perf\left(\frac{ops}{s}\right) = Power(W) \times Efficiency\left(\frac{ops}{joule}\right)$$

W. J. Dally, Keynote IITC 2012



Power Supply (regulation) + Power Consumption + Cooling

Operator_cost + Data_movement_cost + Storage_cost

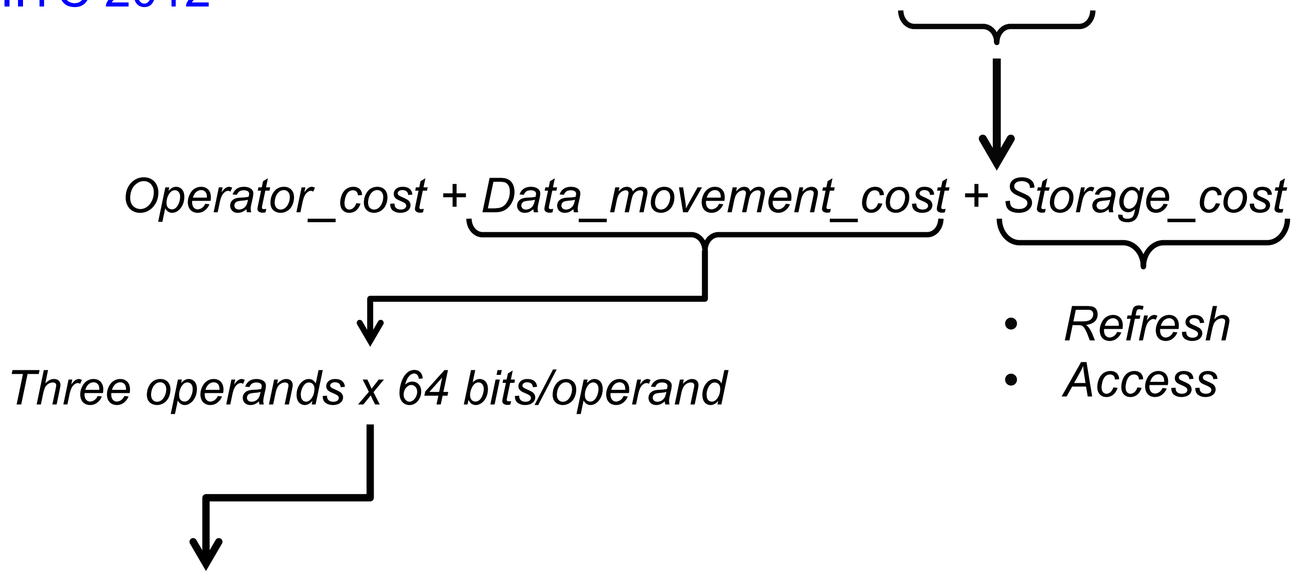
Specialization → *heterogeneity, asymmetry, technology diversity*

*

Energy Cost of Data Management

$$Perf\left(\frac{ops}{s}\right) = Power(W) \times Efficiency\left(\frac{ops}{joule}\right)$$

W. J. Dally, Keynote IITC 2012



Operator_cost + Data_movement_cost + Storage_cost

Three operands x 64 bits/operand

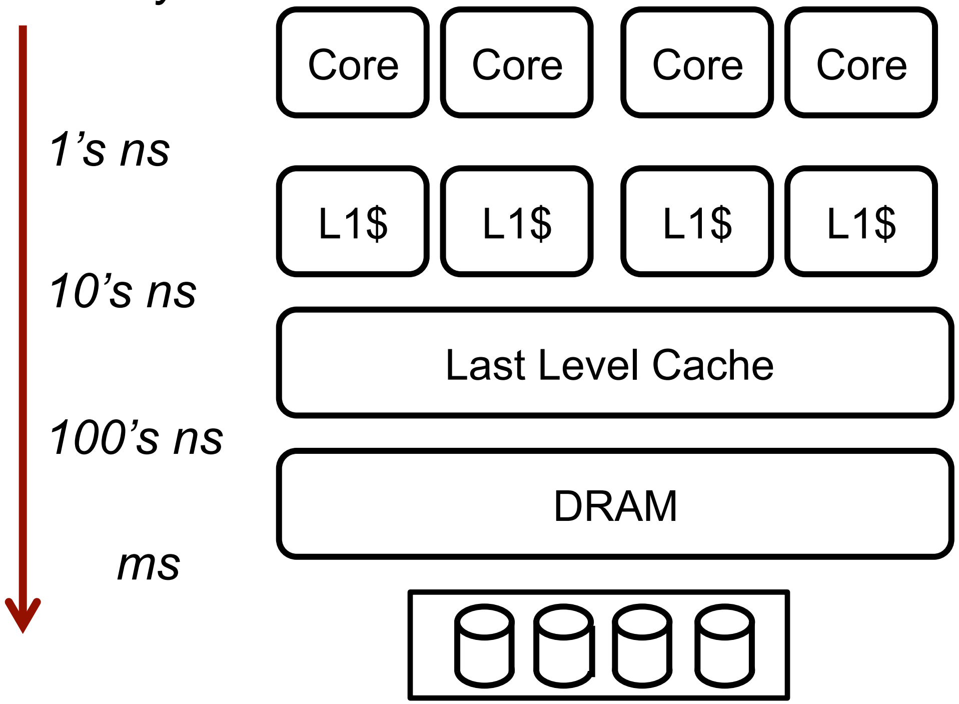
- *Refresh*
- *Access*

$$DataMovementEnergy = \#\,bits \times dist-mm \times energy-bit-mm$$

*S. Borkar and A. Chien, “The Future of Microprocessors, *CACM*, May 2011

Interconnect Energy Taper: Electrical

Data Access Latency


1's ns

10's ns

100's ns

ms

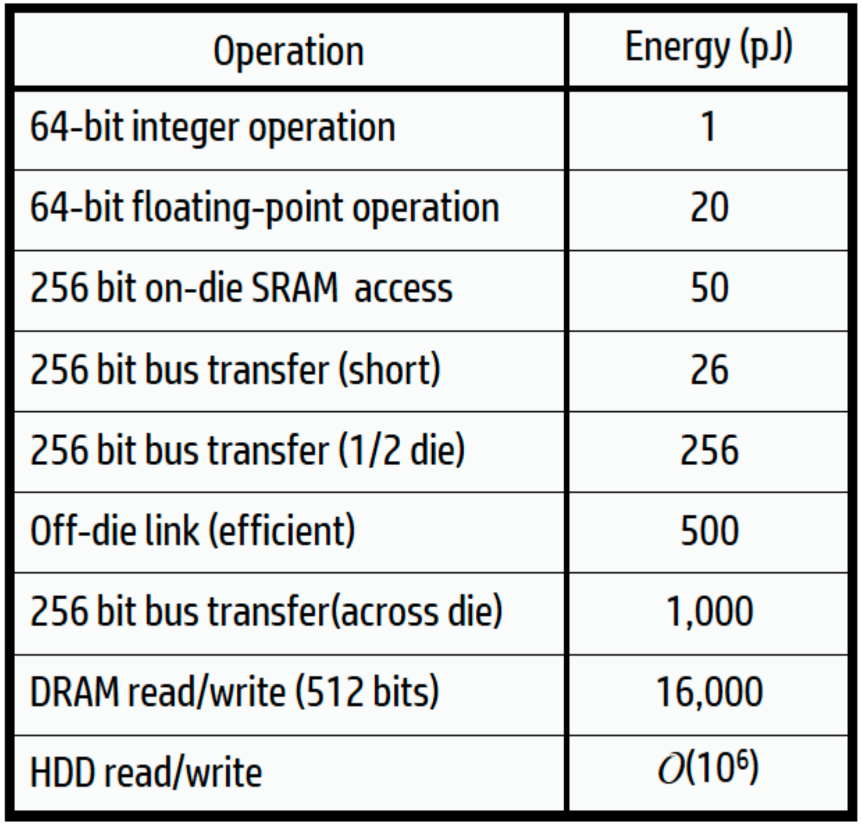
Core Core Core Core

L1$ L1$ L1$ L1$

Last Level Cache

DRAM

Data Access Energy



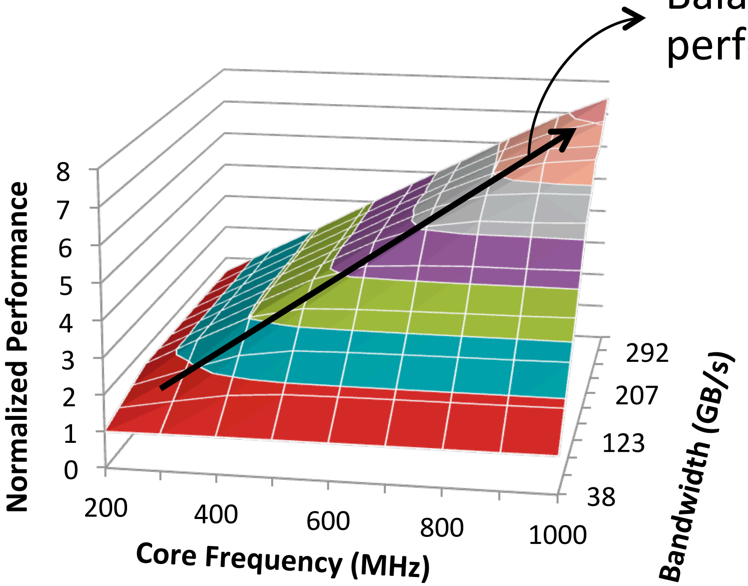
Operation	Energy (pJ)
64-bit integer operation	1
64-bit floating-point operation	20
256 bit on-die SRAM access	50
256 bit bus transfer (short)	26
256 bit bus transfer (1/2 die)	256
Off-die link (efficient)	500
256 bit bus transfer(across die)	1,000
DRAM read/write (512 bits)	16,000
HDD read/write	$\mathcal{O}(10^6)$

28nm CMOS, DDR3

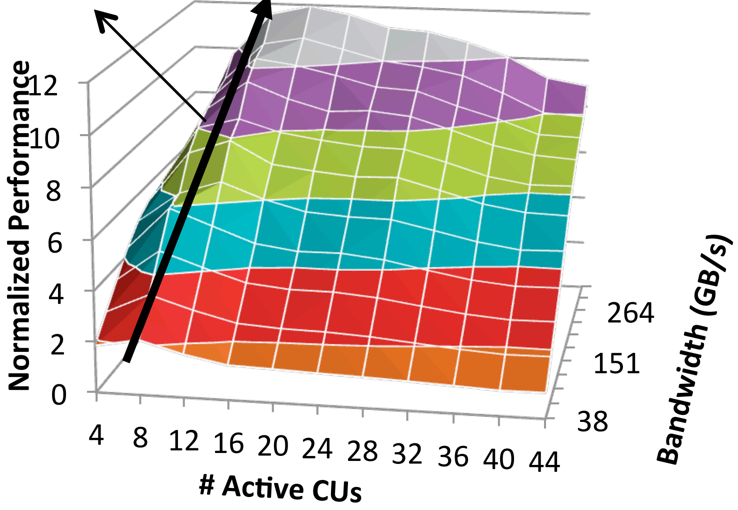
Courtesy Greg Astfalk, HP

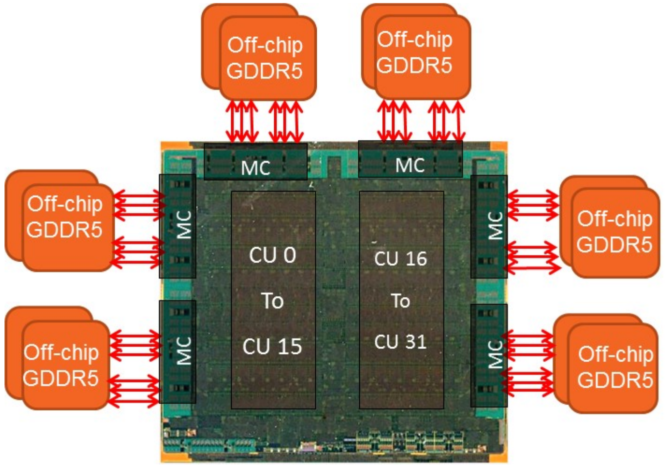
- Relative costs of compute and memory accesses
  - Time and energy costs have shifted to data movement

Shift in the Balance Point



Balance plane for performance and energy



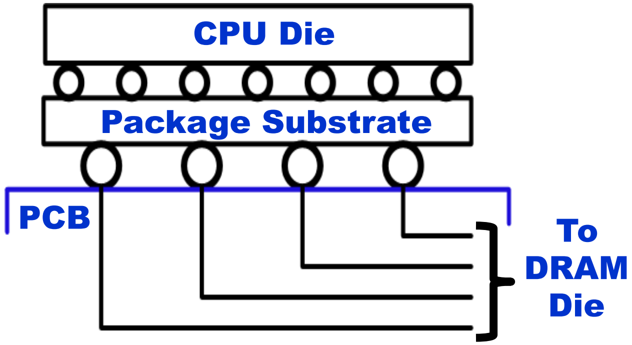


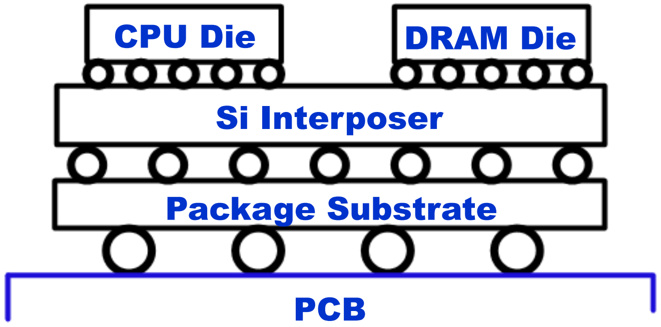
Hardware Balance

Relative energy costs of compute and memory access ⟷ *Relative ops/byte demand of application*

I. Paul, W. Huang, M. Arora, and S. Yalamanchili, "Harmonia: Balancing Compute and Memory Power in High Performance GPUs," *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, June 2015.

Pin Bandwidth Challenges[1]





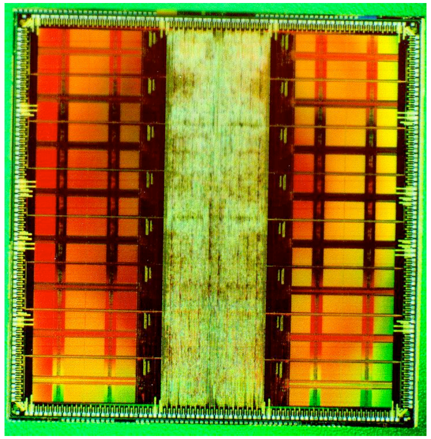
- Number of transistors/die continues to grow
- Number of pins growing at a slower rate than #transistors
- Number of supply pins are crowding out data pins
  - Reducing supply current/pin limits growth of #transistor/die

Data pin bandwidth is not growing as fast as number of transistors on chip

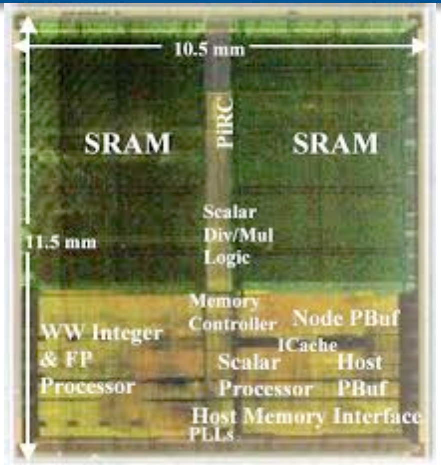
[1] P. Stanley-Marbell, V. C. Cabezas, and R. P. Luijten, " Pinned to the Wall – Impact of Packaging and Applications on the Memory and Power Walls," *IEEE/ACM international symposium on Low-Power Electronics and Design (ISPLED)*, 2011

Re-Emergence of Processing In (Near) Memory

1990's

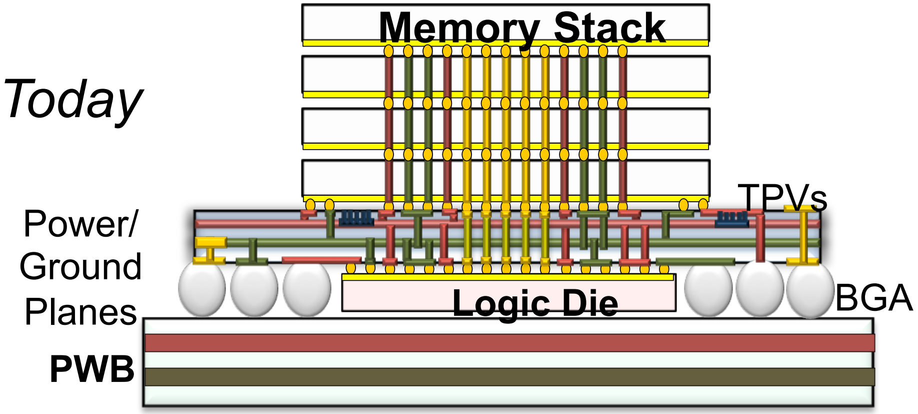


Kogge – Execube (4K DRAM + 100K gate parallel processor (www3.nd.edu)

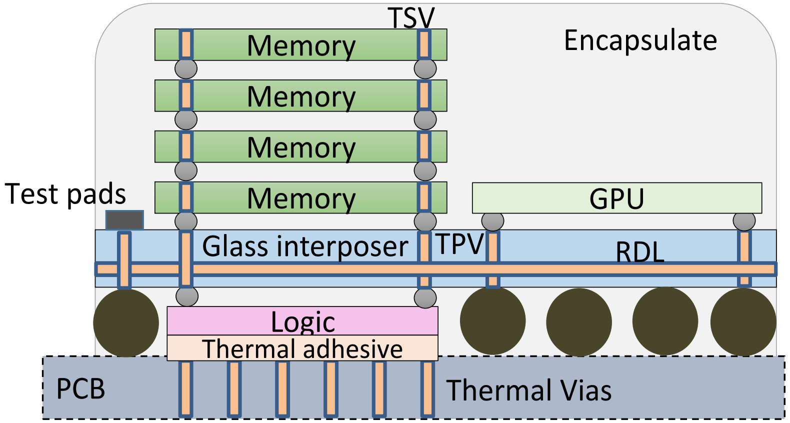


Draper et.al, DIVA chip (isi.edu)

Today



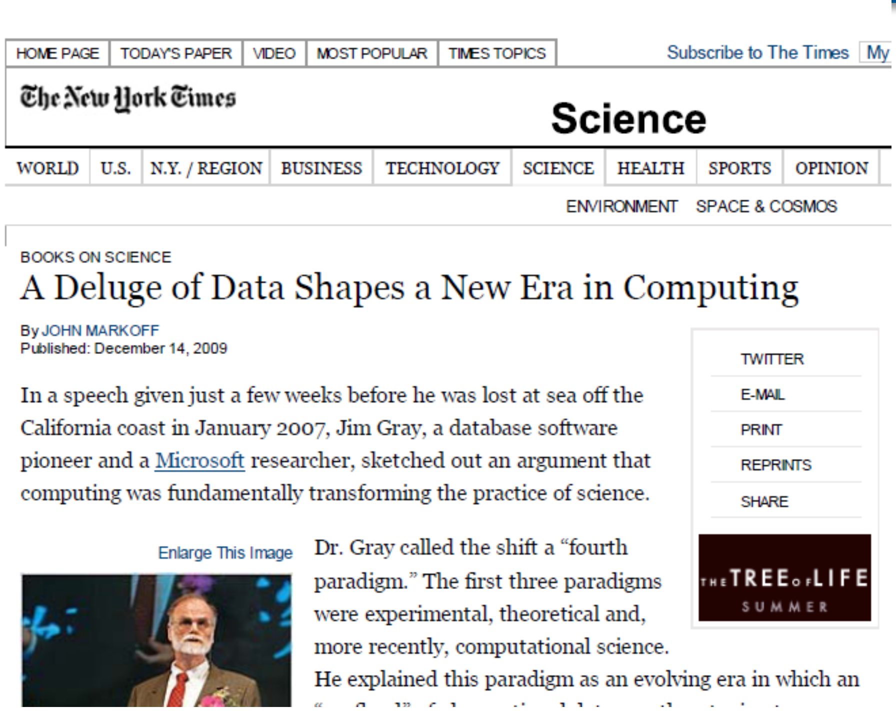
3D Interposer System



Courtesy Gokul Kumar

The Data Tsunami

www.hq.unu.edu



HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS | Subscribe to The Times | My

The New York Times

Science

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

ENVIRONMENT | SPACE & COSMOS

BOOKS ON SCIENCE

A Deluge of Data Shapes a New Era in Computing

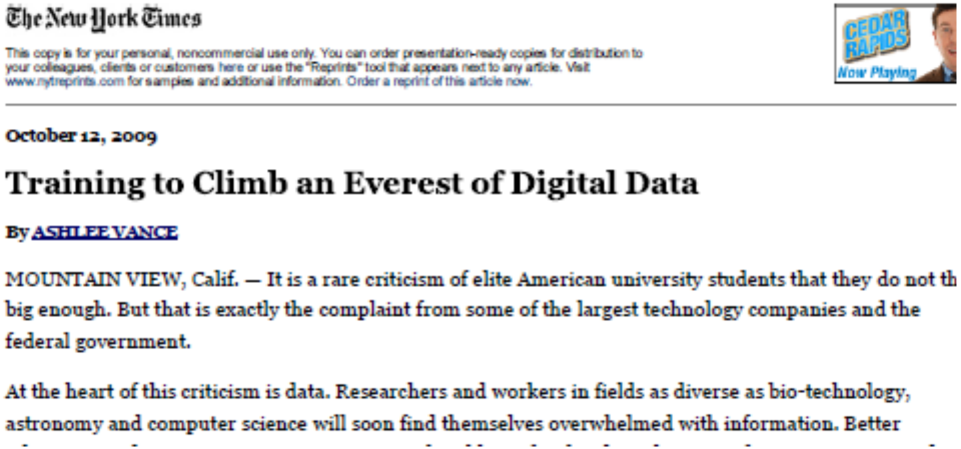
By JOHN MARKOFF
Published: December 14, 2009

TWITTER
E-MAIL
PRINT
REPRINTS
SHARE

In a speech given just a few weeks before he was lost at sea off the California coast in January 2007, Jim Gray, a database software pioneer and a Microsoft researcher, sketched out an argument that computing was fundamentally transforming the practice of science.

Enlarge This Image

Dr. Gray called the shift a "fourth paradigm." The first three paradigms were experimental, theoretical and, more recently, computational science. He explained this paradigm as an evolving era in which an



The New York Times

This copy is for your personal, noncommercial use only. You can order presentation-ready copies for distribution to your colleagues, clients or customers here or use the "Reprints" tool that appears next to any article. Visit www.nytreprints.com for samples and additional information. Order a reprint of this article now.

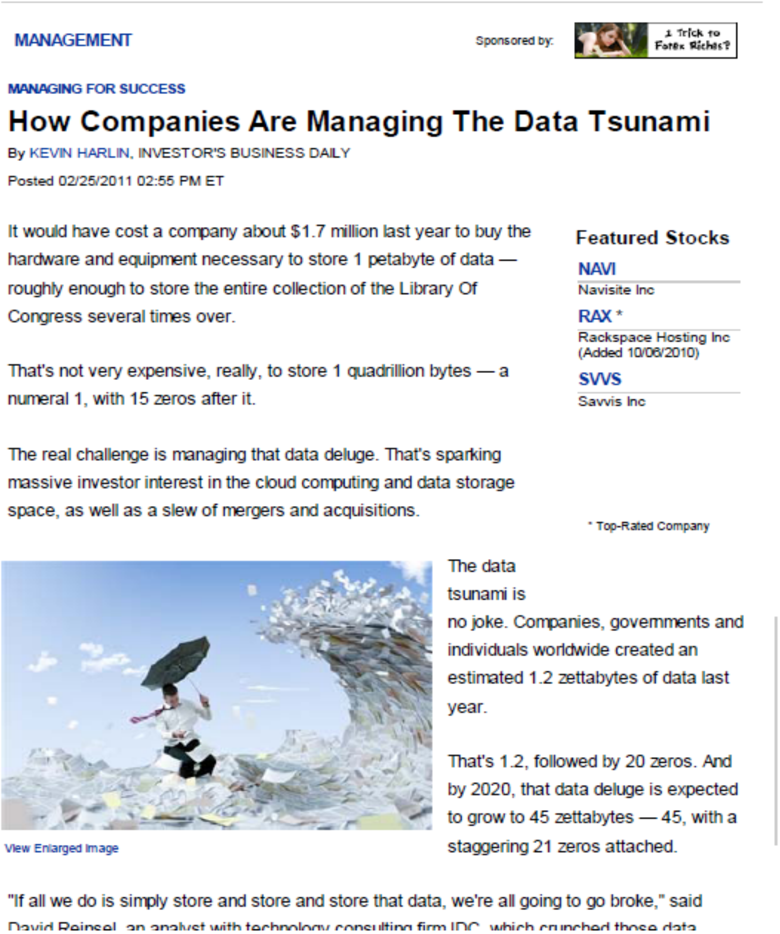
October 12, 2009

Training to Climb an Everest of Digital Data

By ASHLEE VANCE

MOUNTAIN VIEW, Calif. — It is a rare criticism of elite American university students that they do not th big enough. But that is exactly the complaint from some of the largest technology companies and the federal government.

At the heart of this criticism is data. Researchers and workers in fields as diverse as bio-technology, astronomy and computer science will soon find themselves overwhelmed with information. Better



MANAGEMENT

MANAGING FOR SUCCESS

How Companies Are Managing The Data Tsunami

By KEVIN HARLIN, INVESTOR'S BUSINESS DAILY

Posted 02/25/2011 02:55 PM ET

It would have cost a company about $1.7 million last year to buy the hardware and equipment necessary to store 1 petabyte of data — roughly enough to store the entire collection of the Library Of Congress several times over.

That's not very expensive, really, to store 1 quadrillion bytes — a numeral 1, with 15 zeros after it.

The real challenge is managing that data deluge. That's sparking massive investor interest in the cloud computing and data storage space, as well as a slew of mergers and acquisitions.

Featured Stocks

NAVI
Navisite Inc

RAX *
Rackspace Hosting Inc
(Added 10/06/2010)

SVVS
Savvis Inc

* Top-Rated Company

View Enlarged Image

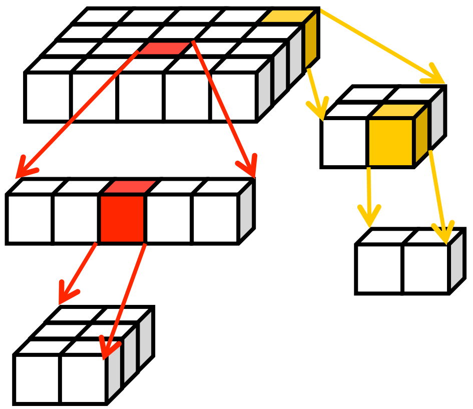
The data tsunami is no joke. Companies, governments and individuals worldwide created an estimated 1.2 zettabytes of data last year.

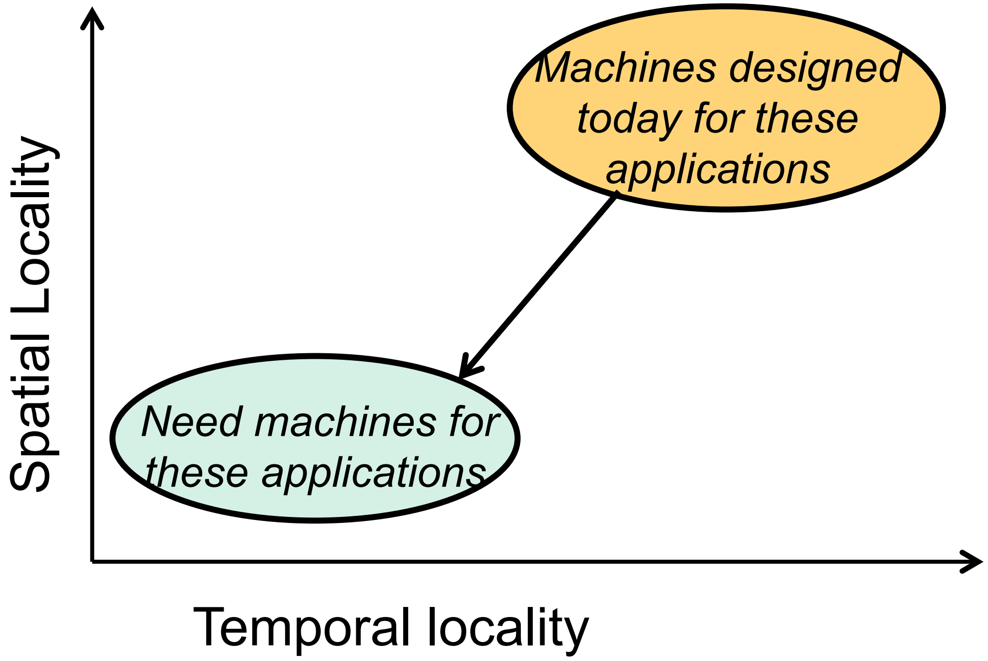
That's 1.2, followed by 20 zeros. And by 2020, that data deluge is expected to grow to 45 zettabytes — 45, with a staggering 21 zeros attached.

"If all we do is simply store and store and store that data, we're all going to go broke," said David Reinsel, an analyst with technology consulting firm IDC, which crunched those data

Shift in Re-Use Patterns: Locality

Adaptive Mesh Refinement (AMR)



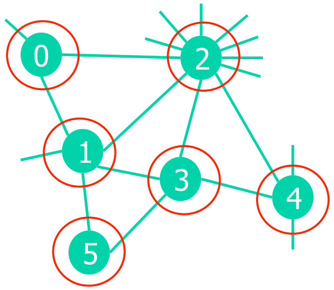


Machines designed today for these applications

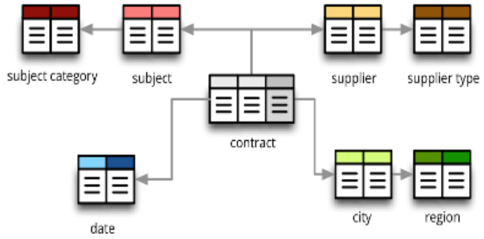
Need machines for these applications

Machine Learning

Graph Analytics



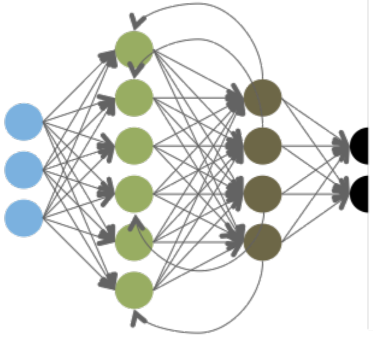
Relational Computing



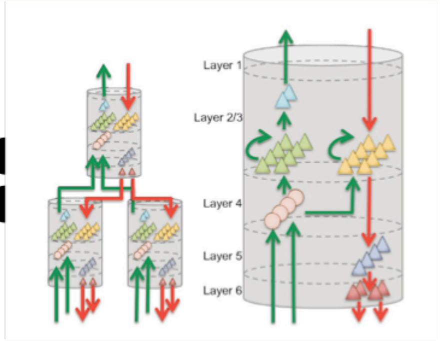
Shift to Finer Arithmetic Density

Neuromorphic

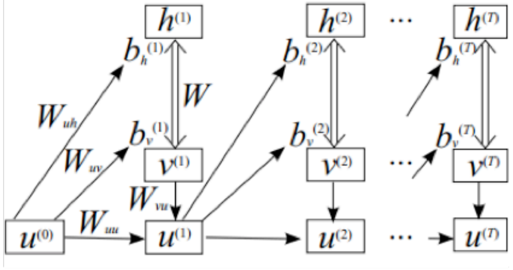
Feedback - Captures temporal history of input sequences



Recurrent Neural Network

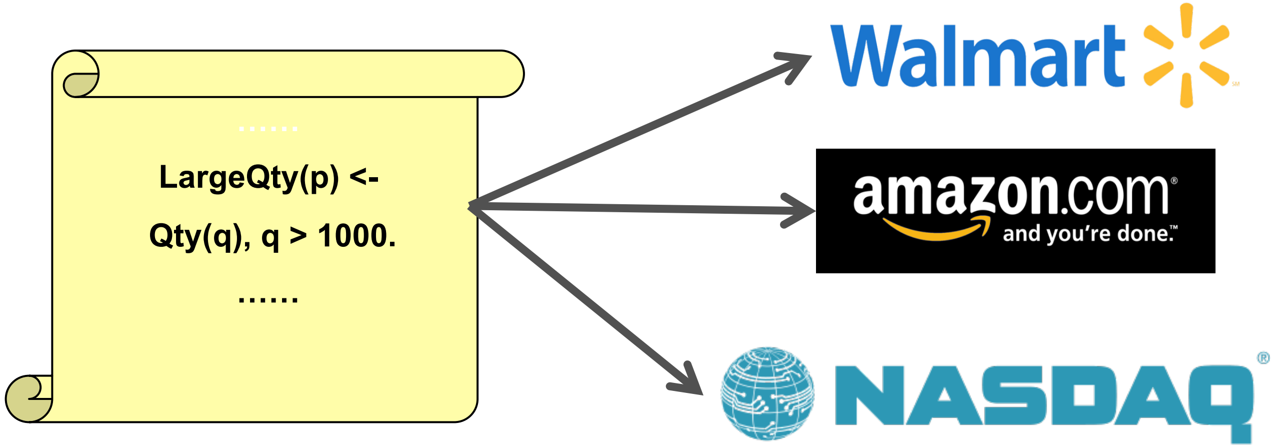


Hierarchical Temporal Memory



Restricted Boltzman Machines (CRBM, TRBM, RTRBM)

Query Procesing



......

LargeQty(p) <-

Qty(q), q > 1000.

......

Walmart

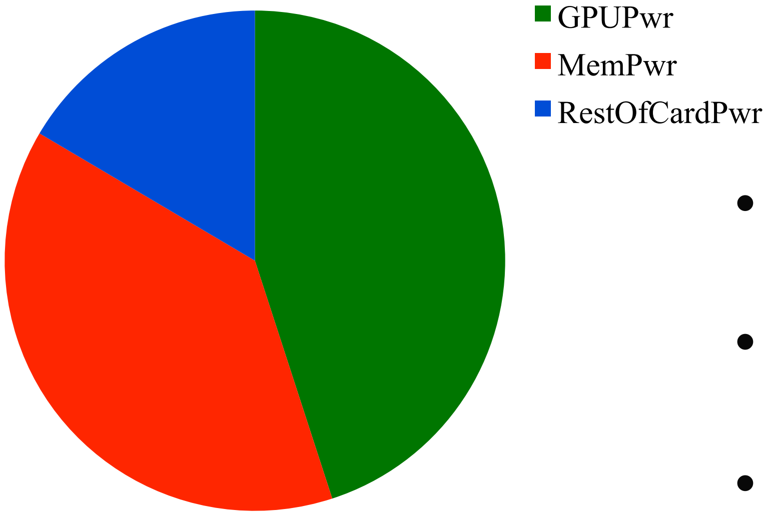
amazon.com and you're done.

NASDAQ

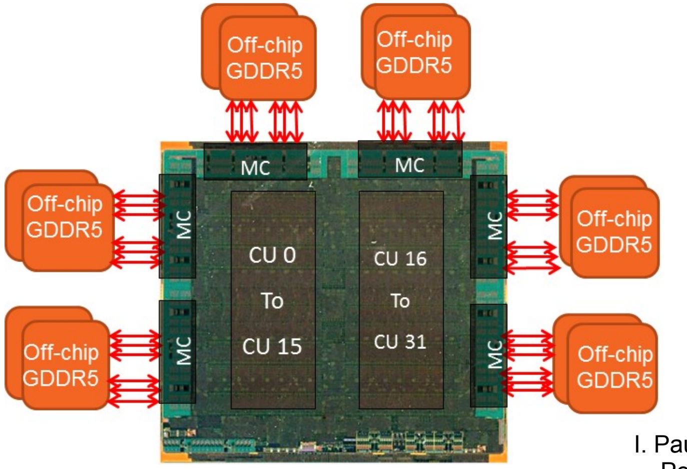
Candidate Application Domains

Relational Computations Over Massive Unstructured Data Sets

Where do the $$ and Energy Go?



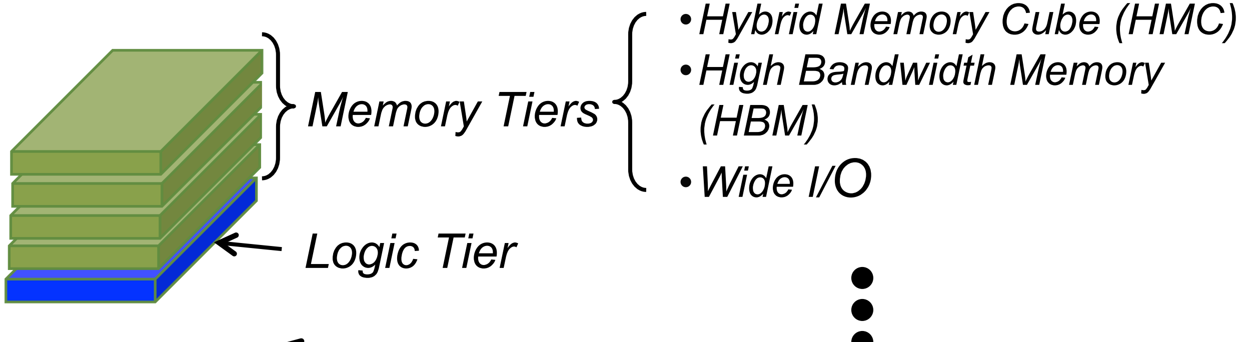
- *Increasing percentage of costs*
- *Increasing percentage of power*
- *Increasing percentage of performance (latency-BW)*
- *Increasing memory intensive applications*



I. Paul, W. Huang, M. Arora, and S. Yalamanchili, “Harmonia: Balancing Compute and Memory Power in High Performance GPUs,” *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, June 2015.

The (Re)-Emergence of Near Data Processing

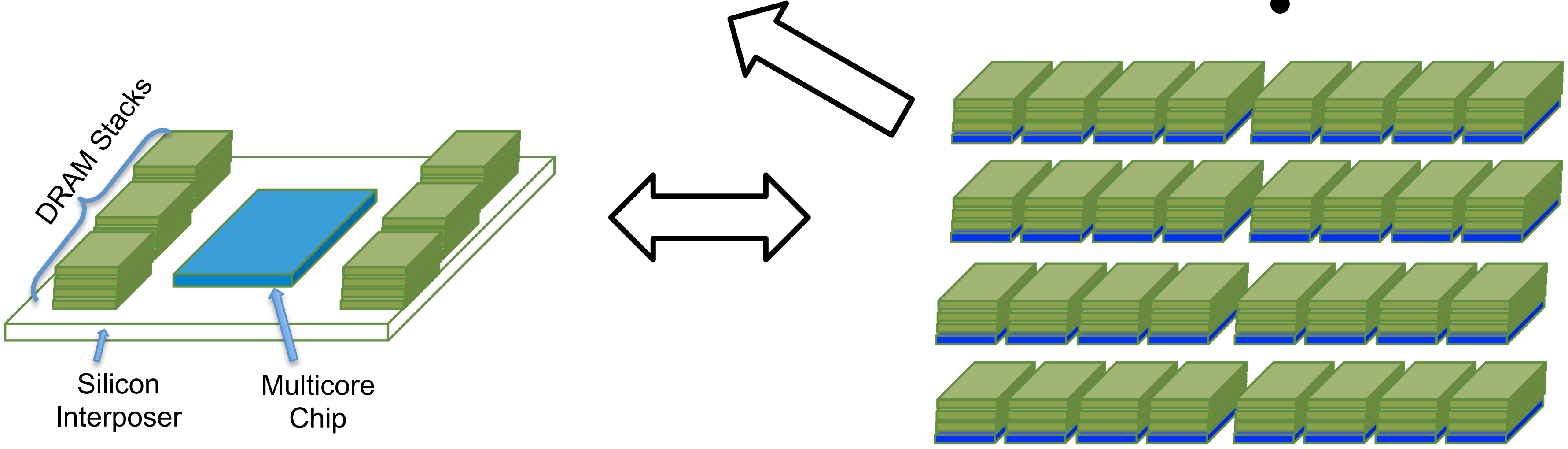
Where are the Networks?



Memory Tiers

- *Hybrid Memory Cube (HMC)*
- *High Bandwidth Memory (HBM)*
- *Wide I/O*

Logic Tier



DRAM Stacks

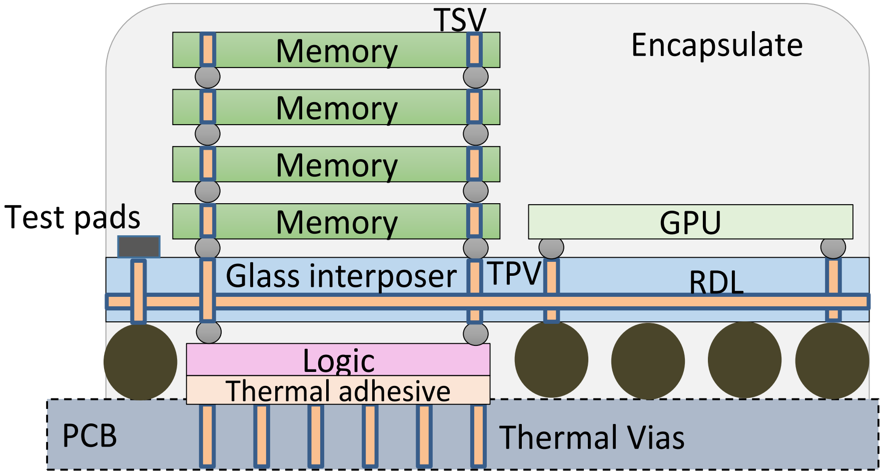
Silicon Interposer

Multicore Chip

Compute Package

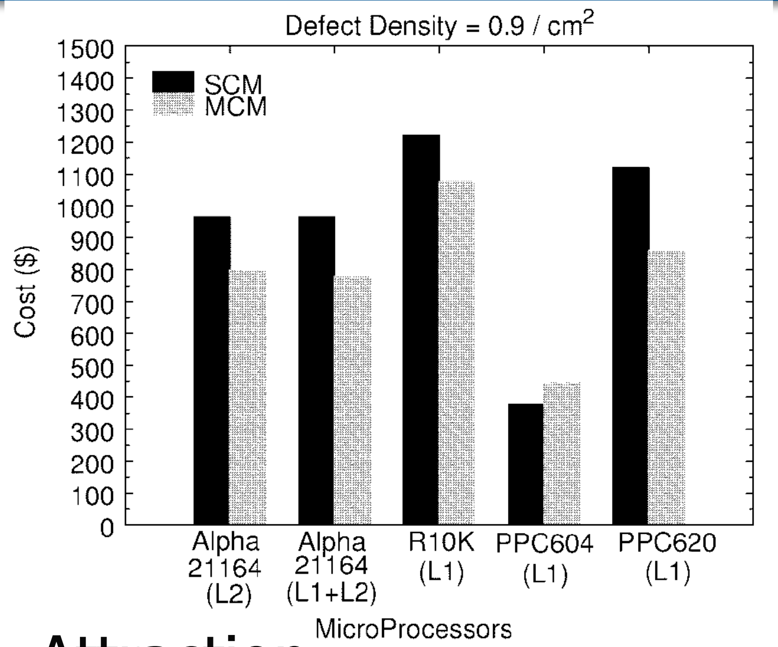
New BW Hierarchy and energy taper

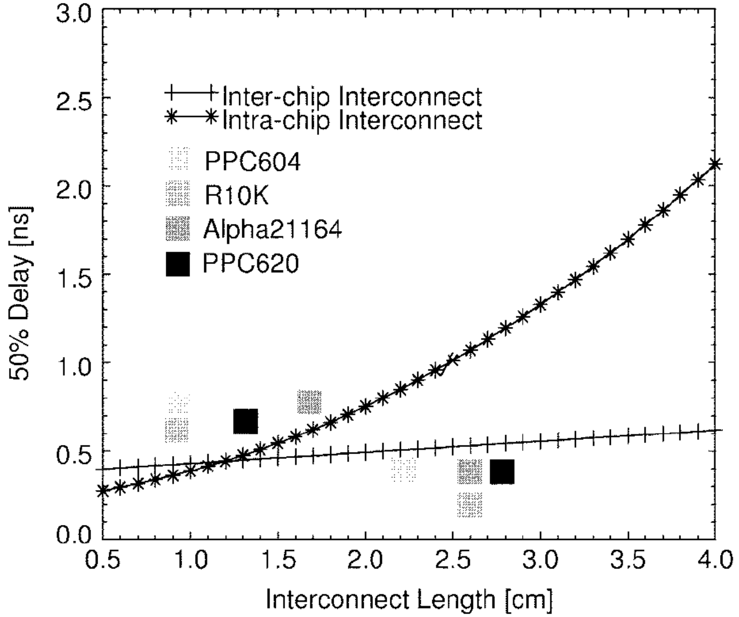
Capacity Tier Memory



Transition to Memory Centric Compute: Inside the Package

Impact of Interposer: Processor-Memory Hierarchy

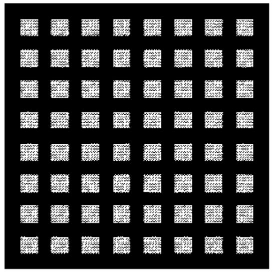




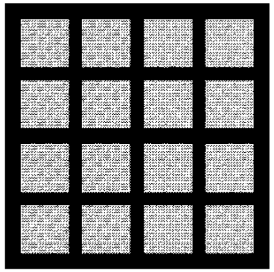
- Attraction
  - Smaller die (better yield), process customization, and larger L2
  - Better thermal behaviors
- Issues
  - Increased die and substrate testing costs (increase in I/Os)
  - Cost

V. Garg, D. Stogner, C. Ulmer, D. E. Schimmel, C. Dislis, S. Yalamanchili, and D. S. Wills, “Early Analysis of Cost/Performance Trade-Offs in MCM Systems,” IEEE Transactions on Components, Packaging, and Manufacturing Technology–Part B, vol. 20, no. 3, August 1997.

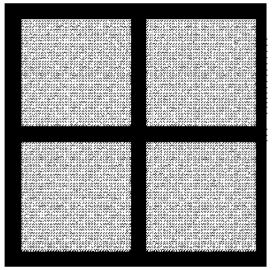
Impact of Interposer: SIMD Image Processor



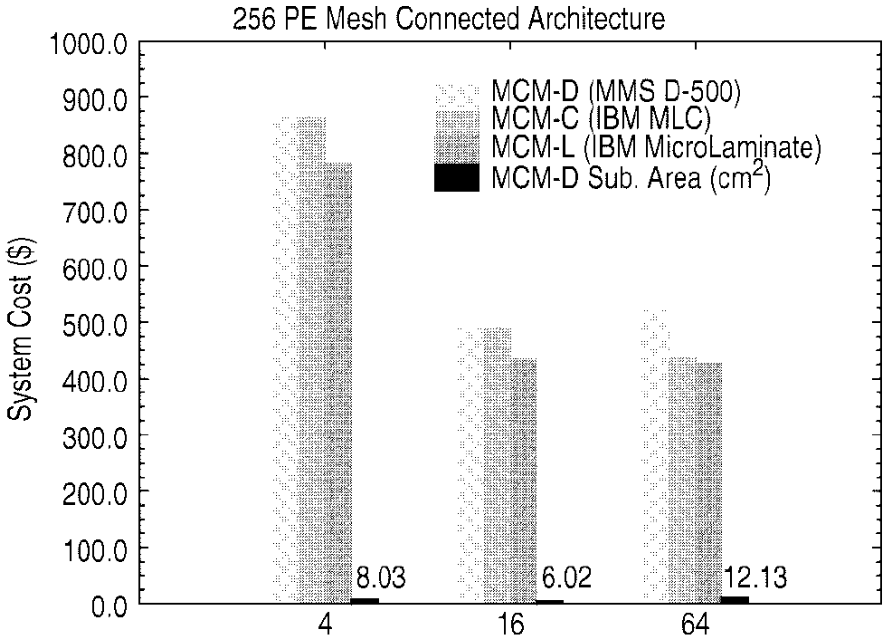
4 PEs/die



16 PEs/die



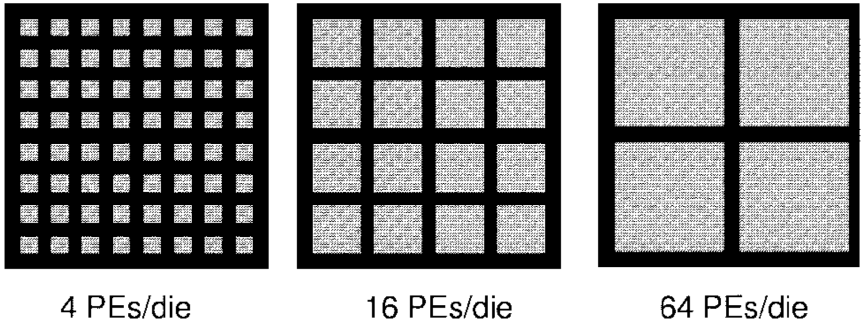
64 PEs/die

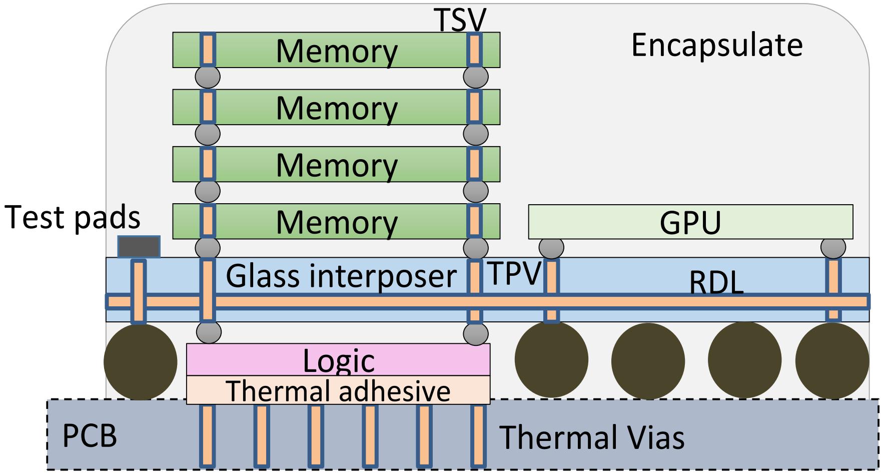


- Trading cost vs. chip size vs number of I/Os
- What is the most cost effective partitioning?

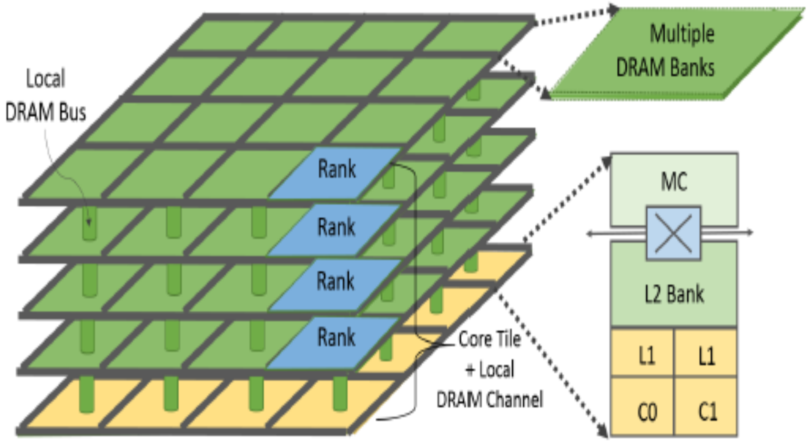
V. Garg, D. Stogner, C. Ulmer, D. E. Schimmel, C. Dislis, S. Yalamanchili, and D. S. Wills, “Early Analysis of Cost/Performance Trade-Offs in MCM Systems,” IEEE Transactions on Components, Packaging, and Manufacturing Technology–Part B, vol. 20, no. 3, August 1997.

The Opportunity: Network in Package



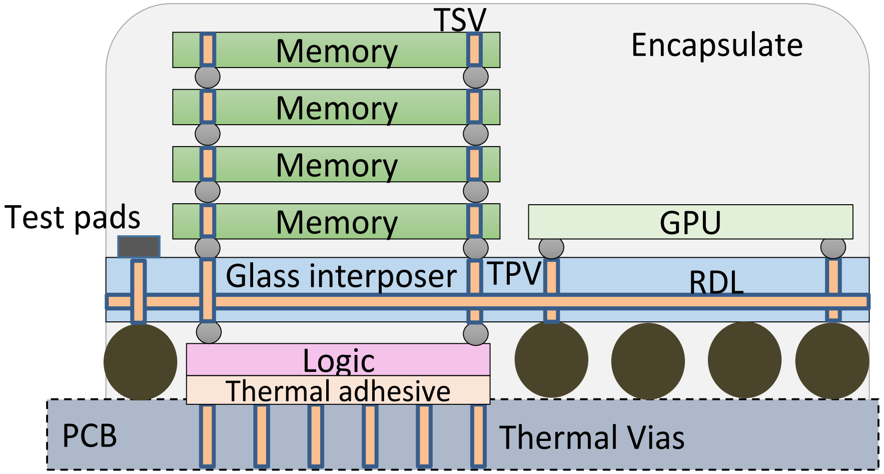


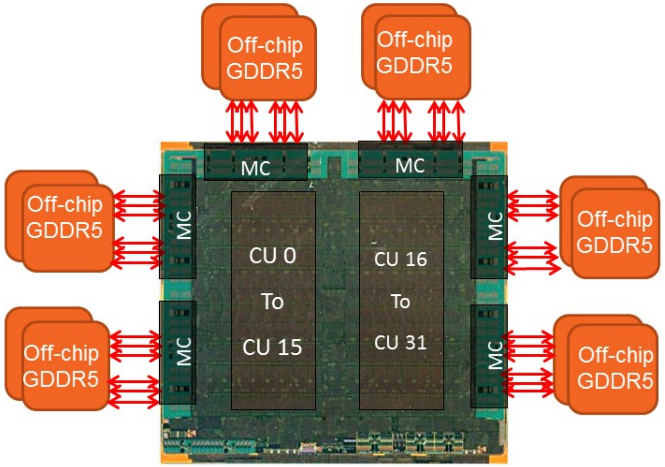
- Smaller micro bumps → increased #connections
- Shorter faster wires in the interposer
- Higher signal integrity → Lower power
- Interposer cost?
- Example: Network on Interposer: Jerger, Kannan, Li, and Loh (MICRO 2014)
- Example: memory networks: G. Kim et. al. PACT 2013

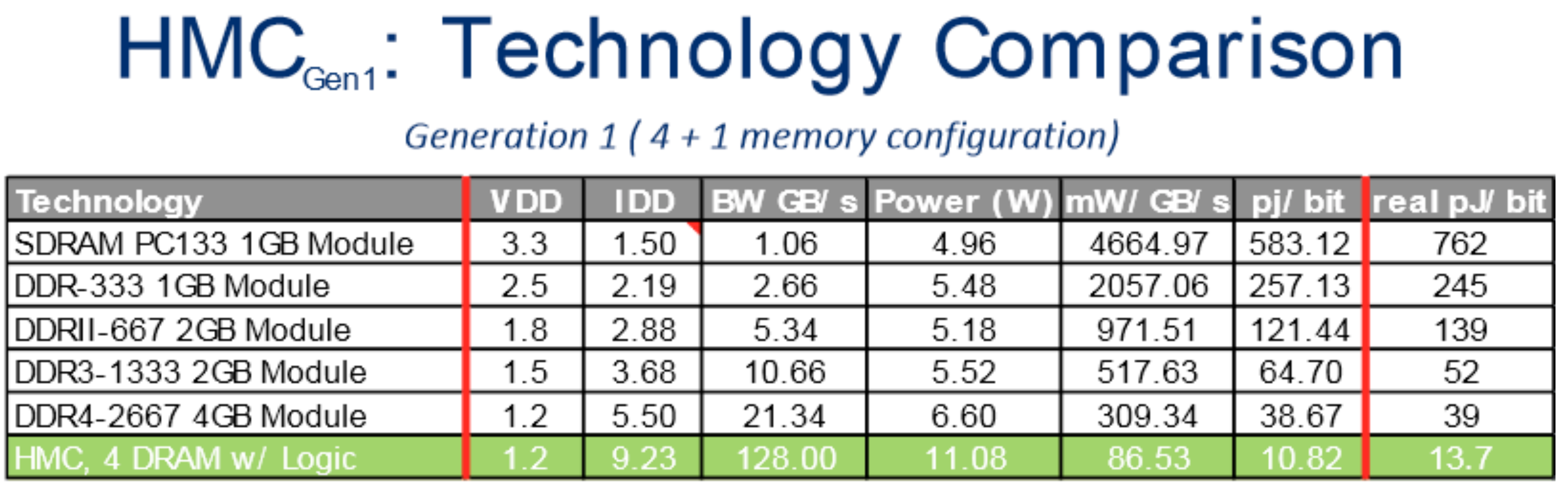


Transition to Memory-Centric Compute: Inside The Stack

3D Multicore Architecture







HMC$_{Gen1}$: Technology Comparison

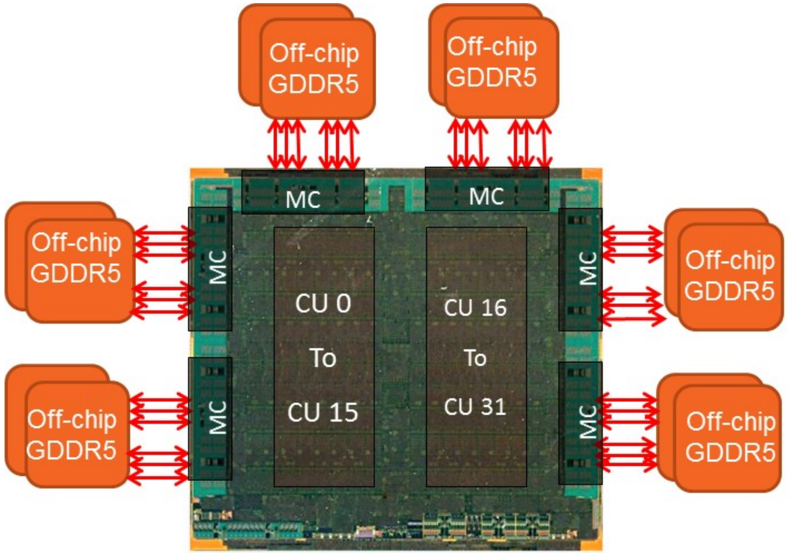
Generation 1 (4 + 1 memory configuration)

Technology	VDD	IDD	BW GB/s	Power (W)	mW/GB/s	pj/bit	real pJ/bit
SDRAM PC133 1GB Module	3.3	1.50	1.06	4.96	4664.97	583.12	762
DDR-333 1GB Module	2.5	2.19	2.66	5.48	2057.06	257.13	245
DDRII-667 2GB Module	1.8	2.88	5.34	5.18	971.51	121.44	139
DDR3-1333 2GB Module	1.5	3.68	10.66	5.52	517.63	64.70	52
DDR4-2667 4GB Module	1.2	5.50	21.34	6.60	309.34	38.67	39
HMC, 4 DRAM w/ Logic	1.2	9.23	128.00	11.08	86.53	10.82	13.7

http://www.extremetech.com/computing/197720-beyond-ddr4-understand-the-differences-between-wide-io-hbm-and-hybrid-memory-cube

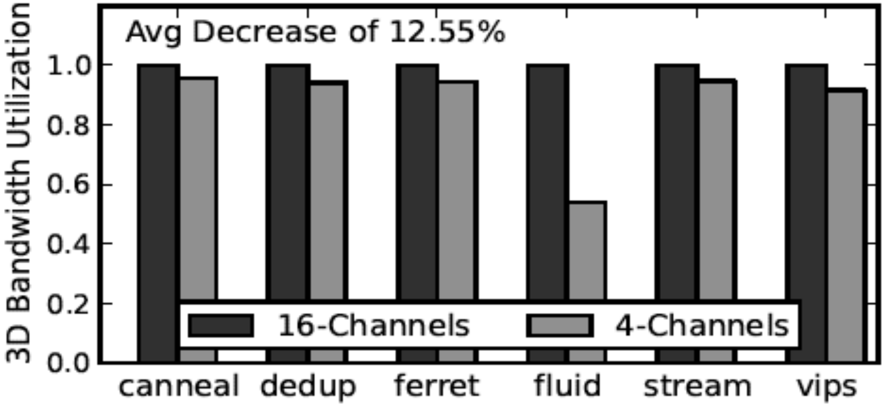
Source of Memory Bandwidth

- Mismatch between bus bandwidth and DRAM access latency
- Over the past two decades density has increased by 1000X and latency reduced by 56% [source:hynix]

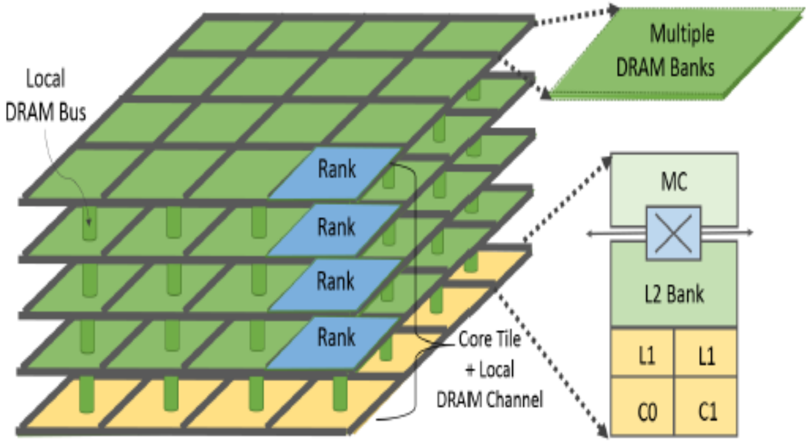


- **Solution**: Have more, narrower channels and exploit parallelism

Parallelism in the Memory System



4 channels with 256 bit bus vs. 16 channels with 64 bit bus



32, x86 cores, Hybrid Memory Cube (HMC) Model

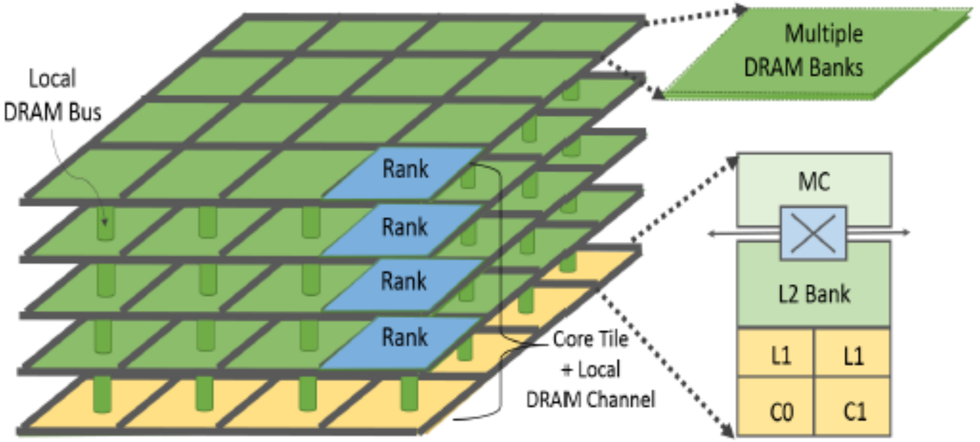
- Move towards more narrower channels
- Increases data cycles improving efficiency and utilization

You can buy bandwidth but you cannot bribe God!
-- Unknown

2D Bandwidth Still Matters!

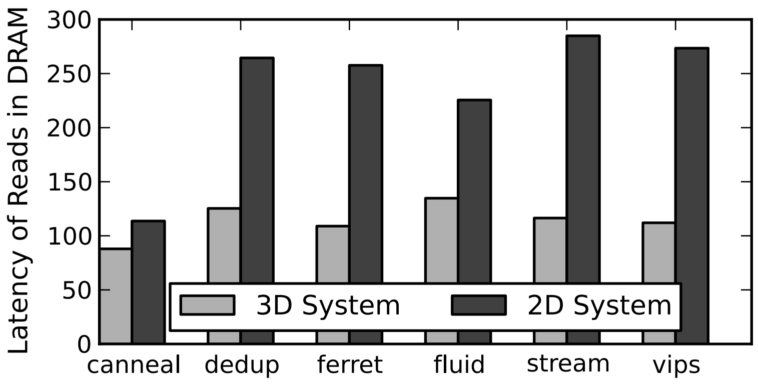
Network Impact of Memory Parallelism

- Tiled 3D memory with 16 channels, similar to HMC
- Distributed directory based coherence with shared L2 banks
- DRAM latency vs. MC queuing

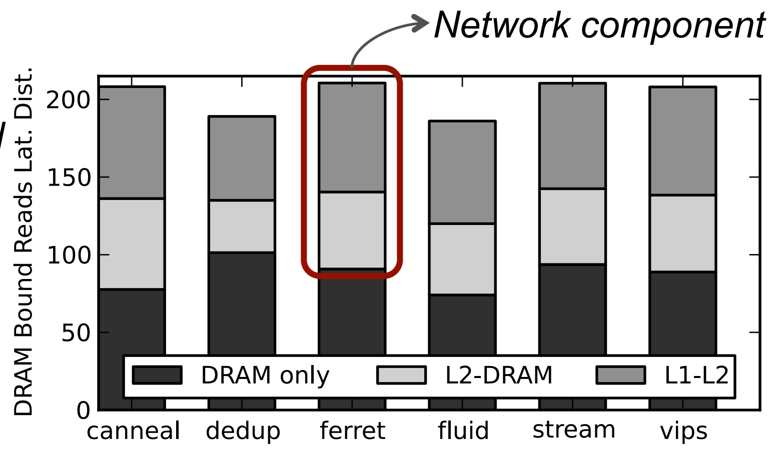


More channels → less load per channel → reduced queueing

2D torus network on compute Tier

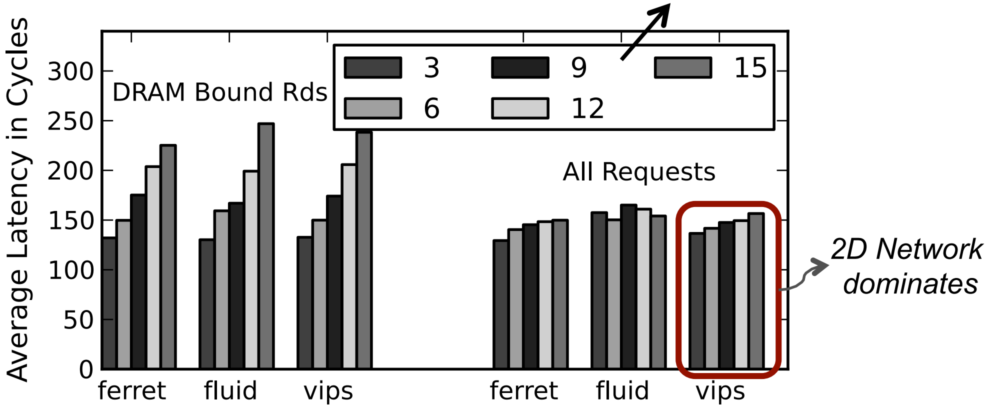


Impact of reduced queuing delays

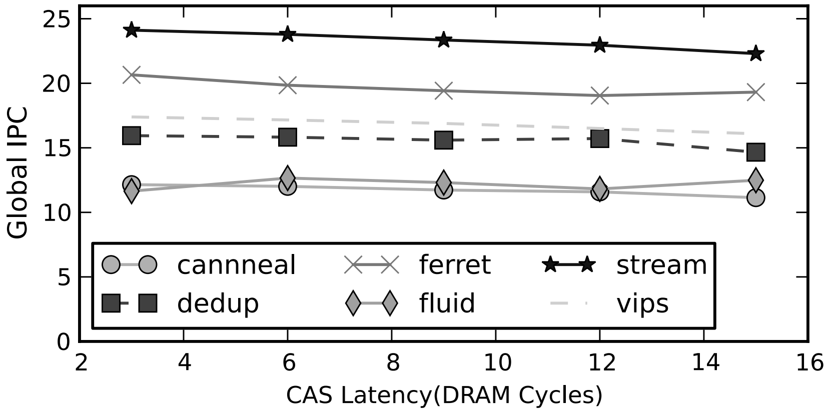


Impact of DRAM Latency vs. Network Latency

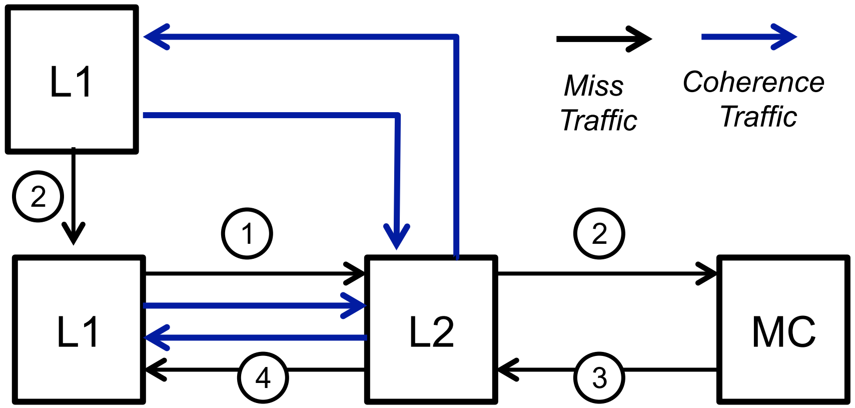
Normalized increase in CAS latency



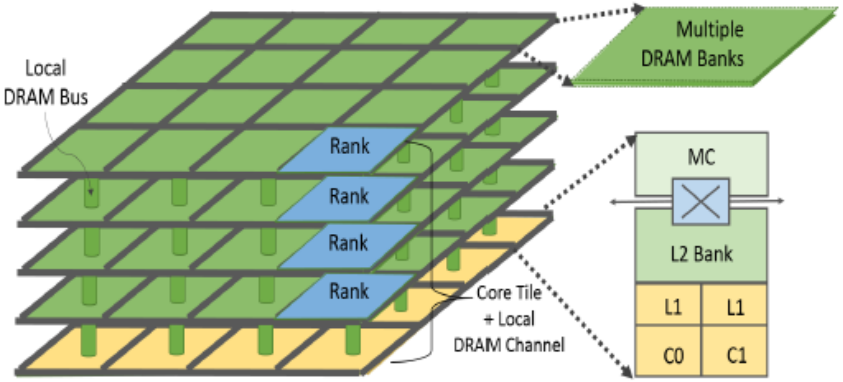
32 cores, 16 memory channels

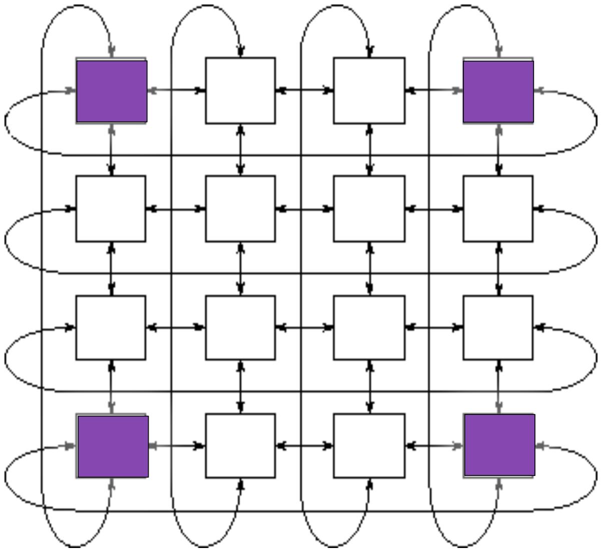


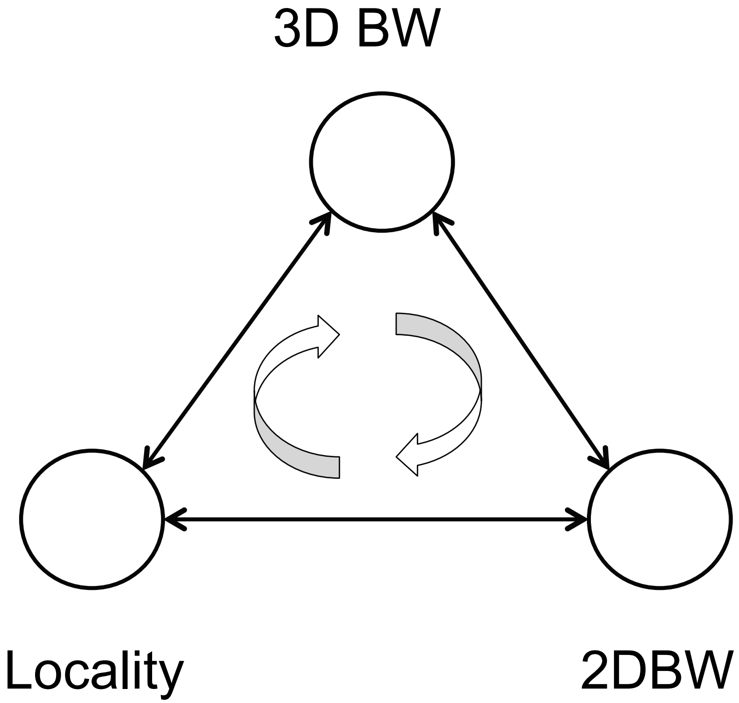
- Impact on all requests is low.
- Coherence requires a request to take multiple hops before being satisfied
- It's the network!



The Opportunity



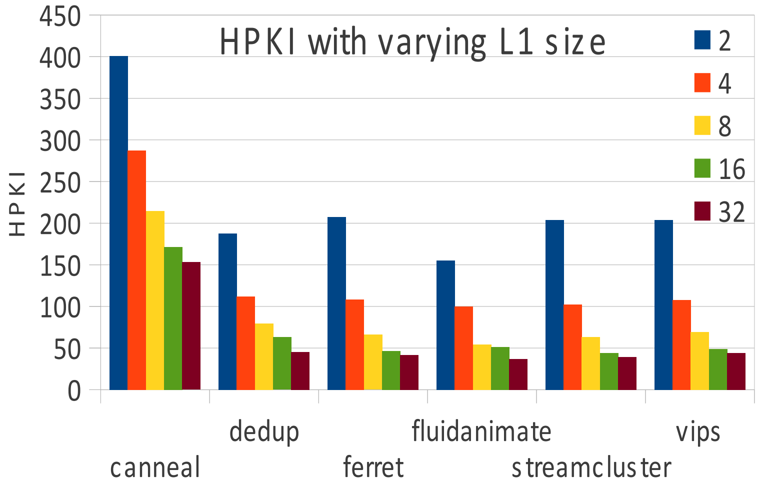


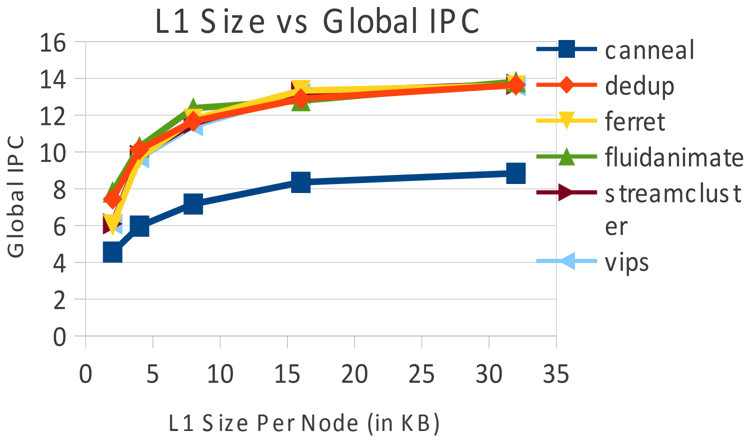


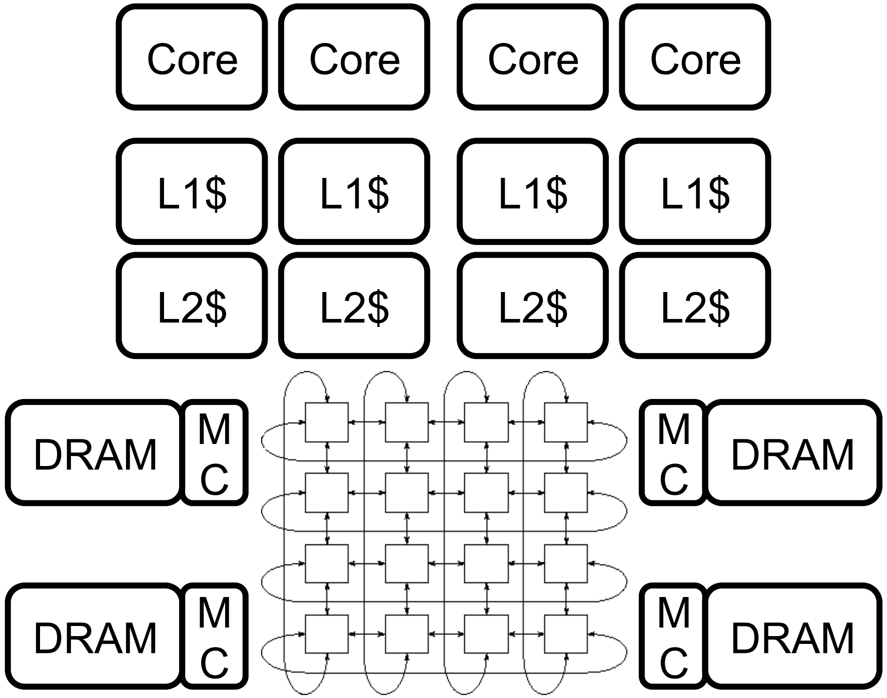
Refactoring the Memory Hierarchy

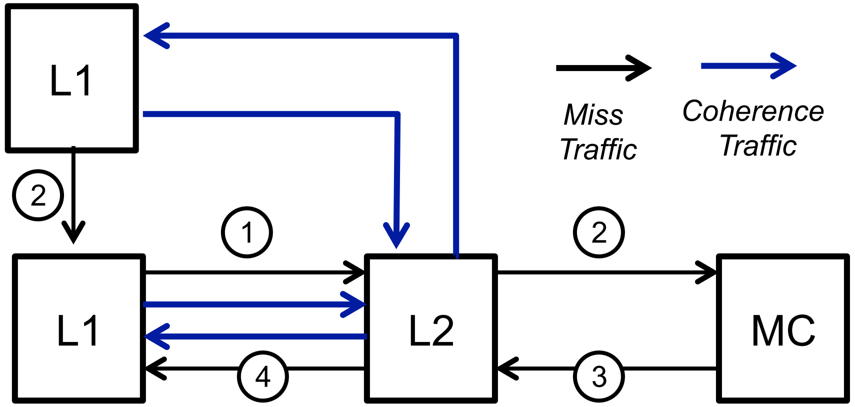
How is the Network Used?

***H**op Count **P**er **K**ilo Instructions (HPKI)*

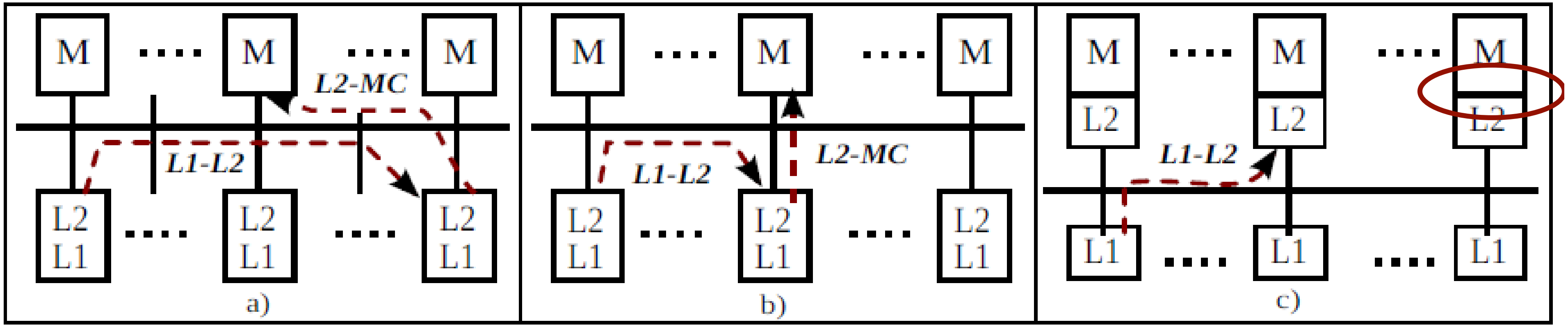






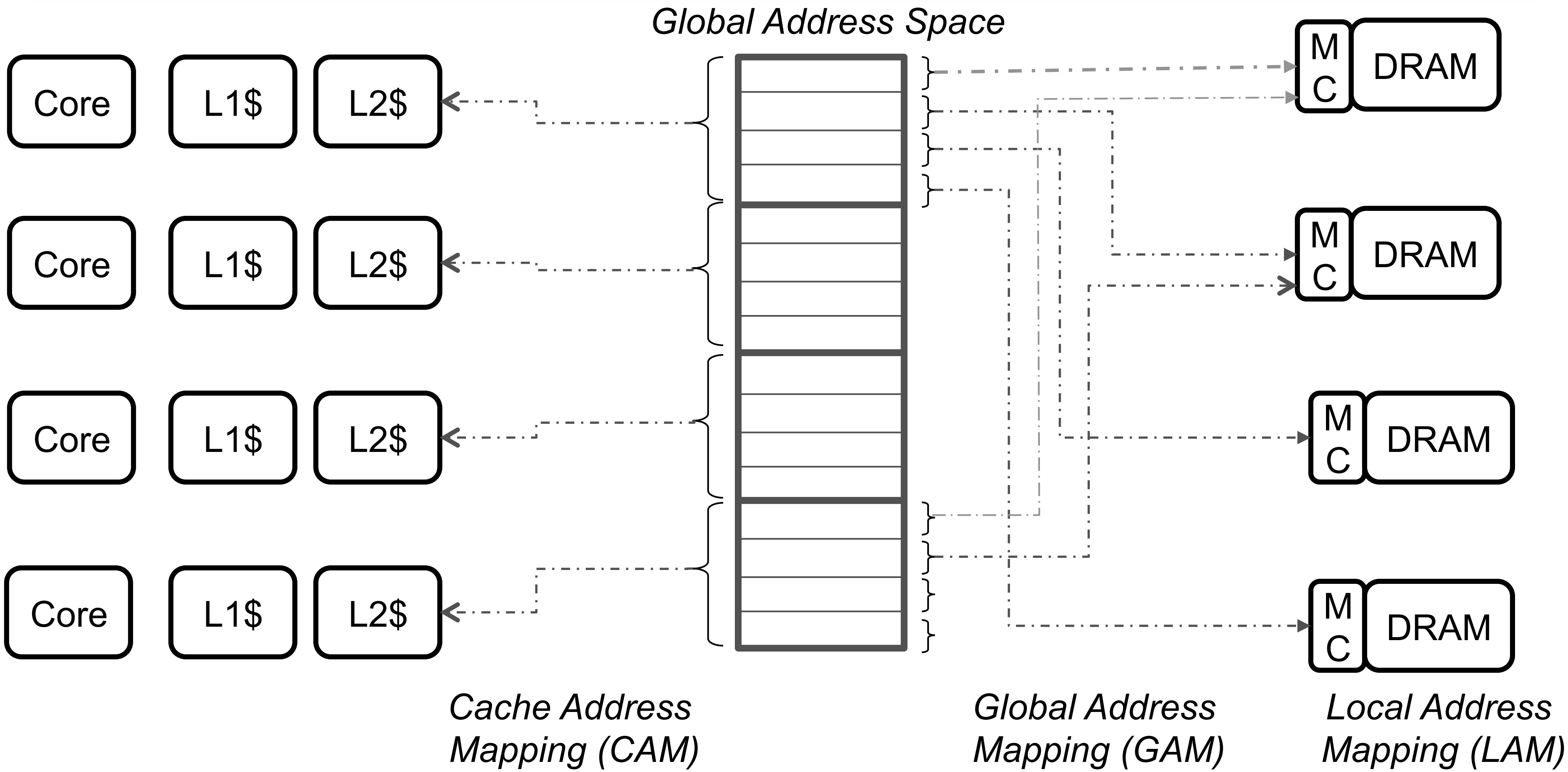


Optimization: Memory Side Caching



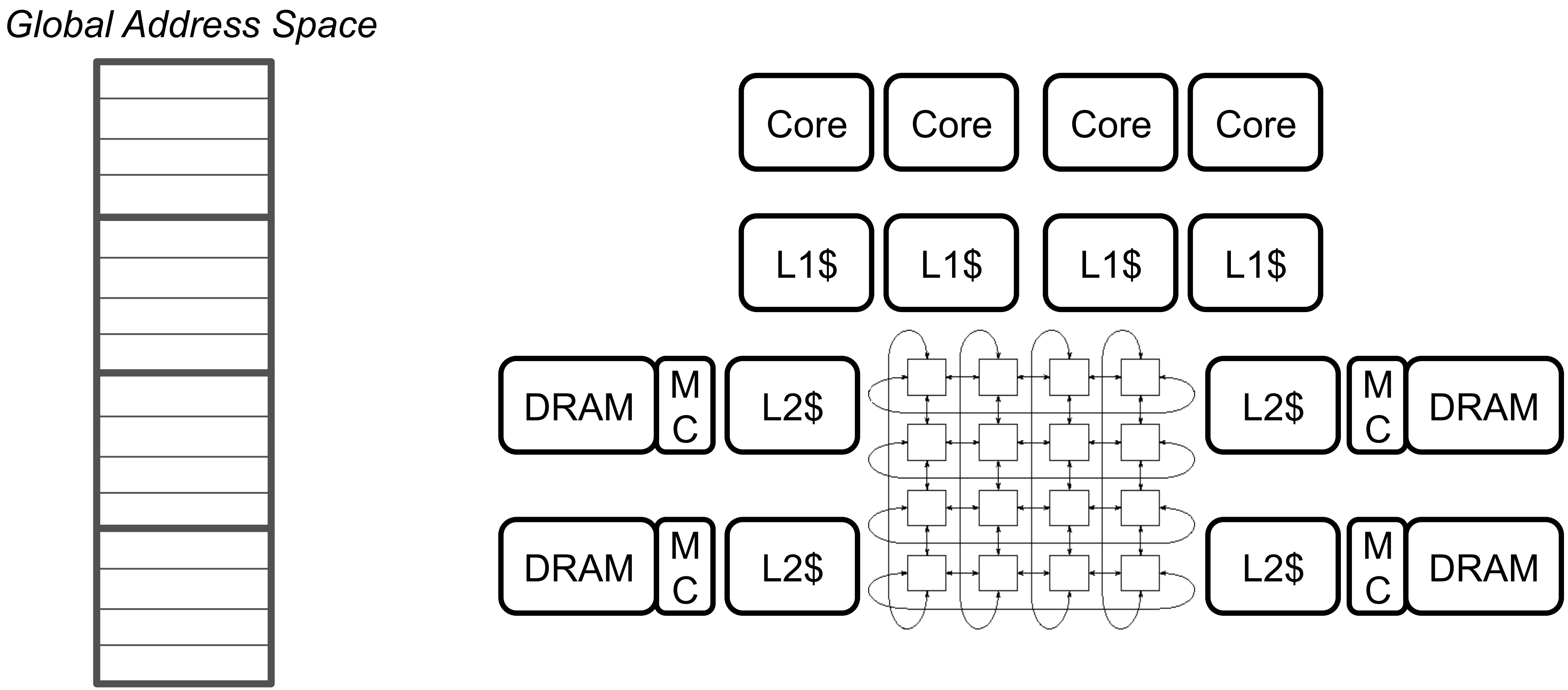
- Refactor memory hierarchy to reduce hop count
- Modify address space mappings to retain/improve locality
- Maximize L2-DRAM BW
  - Remove serialization latency

Importance of Address Space Management: Locality



- Different mapping functions determine parallelism in memory and network traffic
- More parallelism → better 3D bandwidth utilization but more load on the network

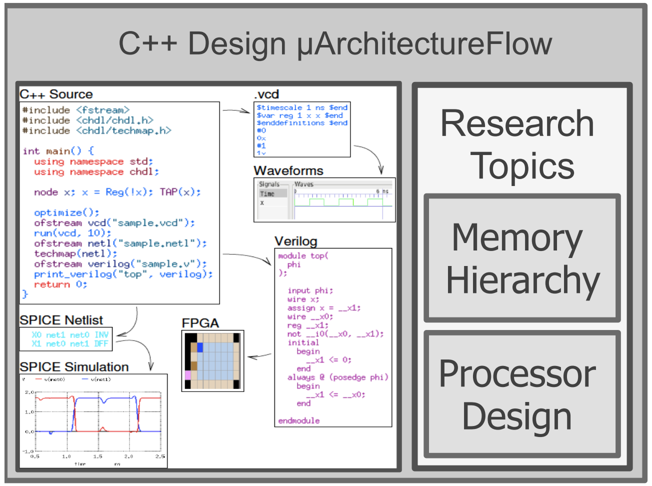
Emphasis on Co-Design



- Refactor the memory hierarchy
- Co-design address space assignment across levels of the memory hierarchy
- Diversity of interconnect technologies

Cymric: A Near Data Processing Architecture

H. Kim, S. Mukhopadhyay, and S. Yalamanchili



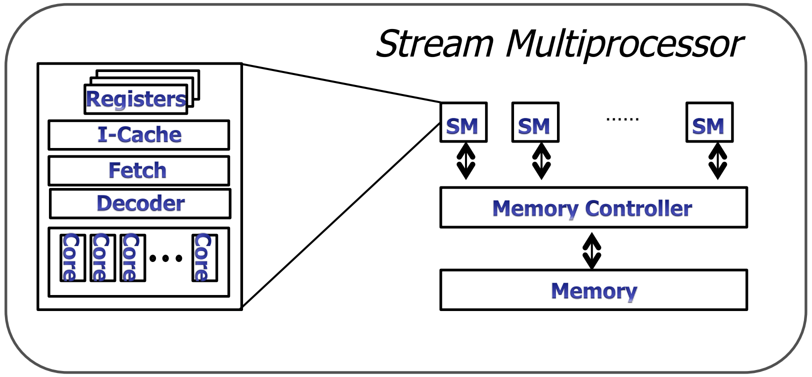
```
#include <fstream>
#include <chdl/chdl.h>
#include <chdl/techmap.h>

int main() {
  using namespace std;
  using namespace chdl;

  node x; x = Reg(!x); TAP(x);

  optimize();
  ofstream vcd("sample.vcd");
  run(vcd, 10);
  ofstream netl("sample.netl");
  techmap(netl);
  ofstream verilog("sample.v");
  print_verilog("top", verilog);
  return 0;
}
```

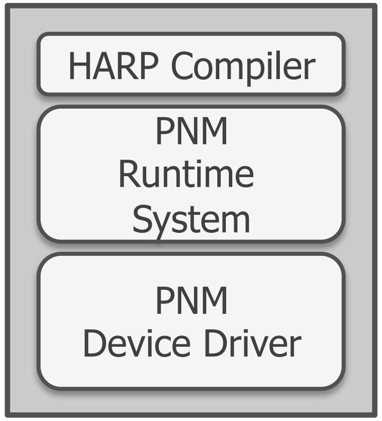
```
$timescale 1 ns $end
$var reg 1 x x $end
$enddefinitions $end
#0
0x
#1
1x
```

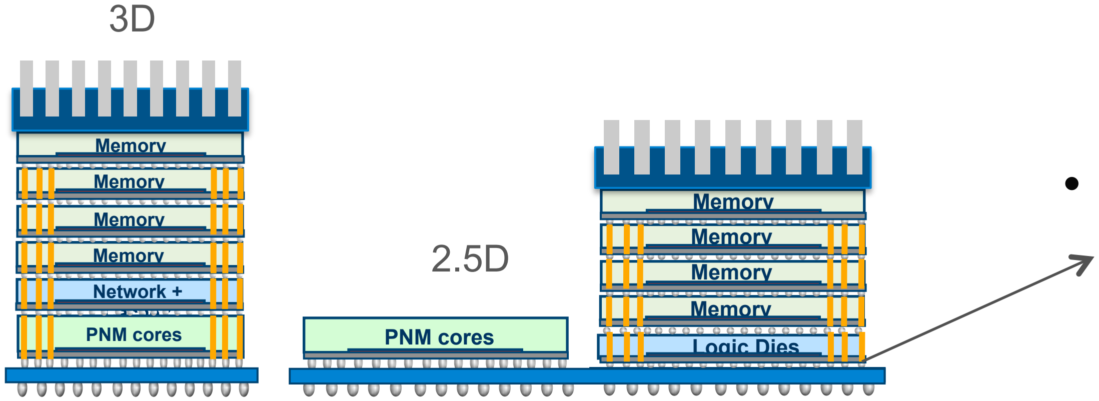
```
module top(
  phi
);

  input phi;
  wire x;
  assign x = __x1;
  wire __x0;
  reg __x1;
  not __i0(__x0, __x1);
  initial
    begin
      __x1 <= 0;
    end
  always @ (posedge phi)
    begin
      __x1 <= __x0;
    end

endmodule
```

```
X0 net1 net0 INV
X1 net0 net1 DFF
```

Custom Lightweight Parametric GPU

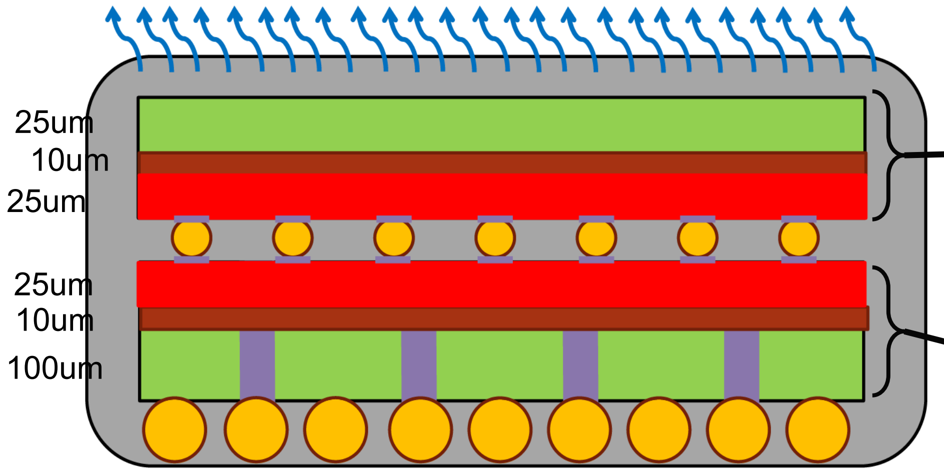






- *High Radix, Small Buffer Networks*
  - *GHC, Slimfly, FB*

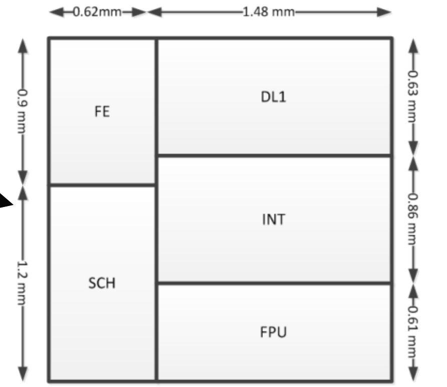
Short Stack: Physical Structure



🟫 Active Layer 🟥 BEOL 🟩 Silicon 🟪 TSV

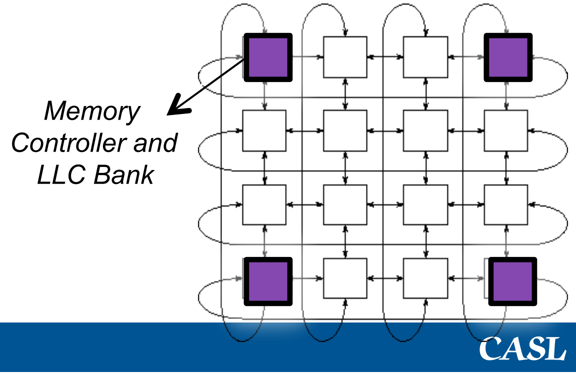
Die Dimension: 8.4 mm X 8.4 mm

LLC & Memory controllers
L2 (per core): 2MB, 4096 sets, 128B, 35 cycles;



16 homogeneous OOO x86 cores
3GHz, 1.0V, max temp 100°C
DL1: 128KB, 4096 sets, 64B
L1: 32KB, 128 sets, 64B, 1 cycles;

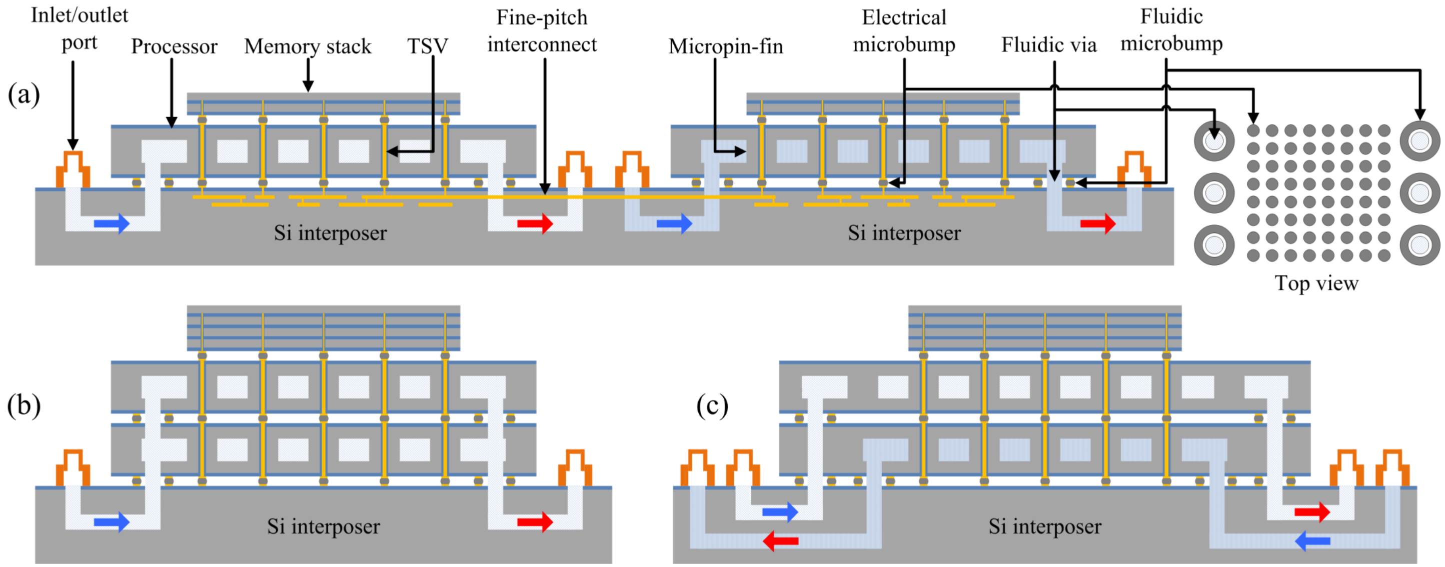
Short Stack



Why is Temperature A NoC Problem?

Thermal Challenges and Microfluidic Cooling

Courtesy Professor Muhannad Bakir (GT/ECE)



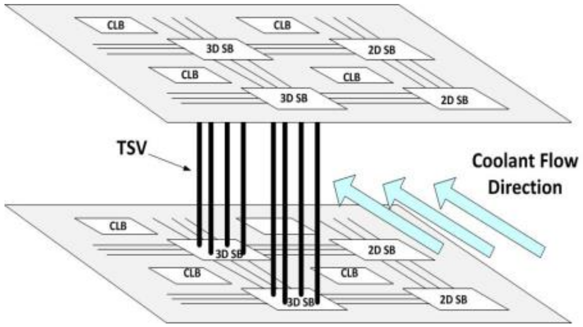
- Fluid flow through the microchannels carry heat out to an external heat exchanger (e.g., heat sink)

This research supported by the DARPA ICECOOL Program

Georgia Institute of Technology ALTERA. UNIVERSITY OF MARYLAND FIU FLORIDA INTERNATIONAL UNIVERSITY DARPA

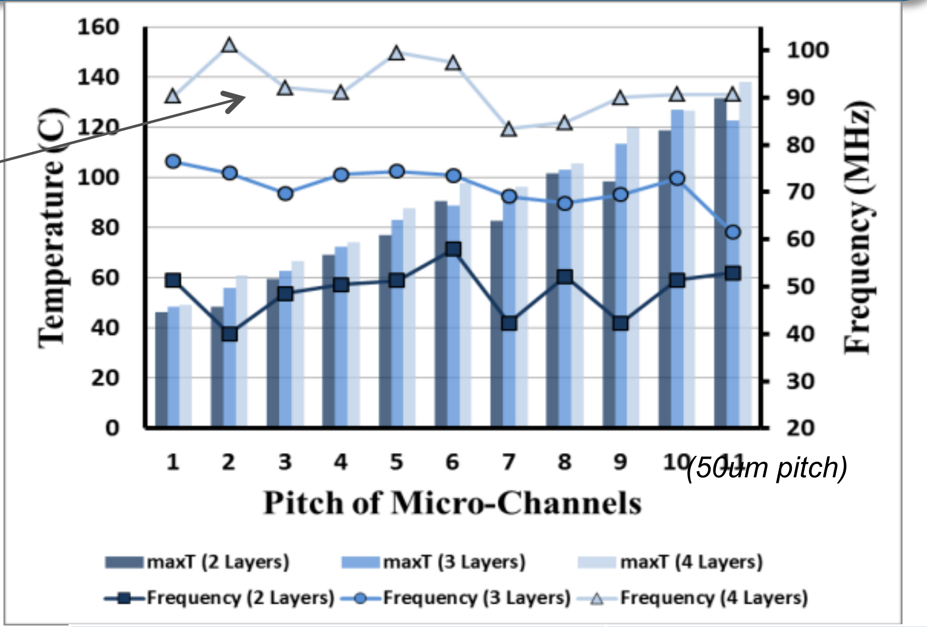
3D FPGA: 3D Bandwidth – Performance Tradeoff

Courtesy: A. Srivastava (UMD)



Physical Architecture of Micro-Fluidically Cooled 3D FPGA

2D vs. 3D congestion



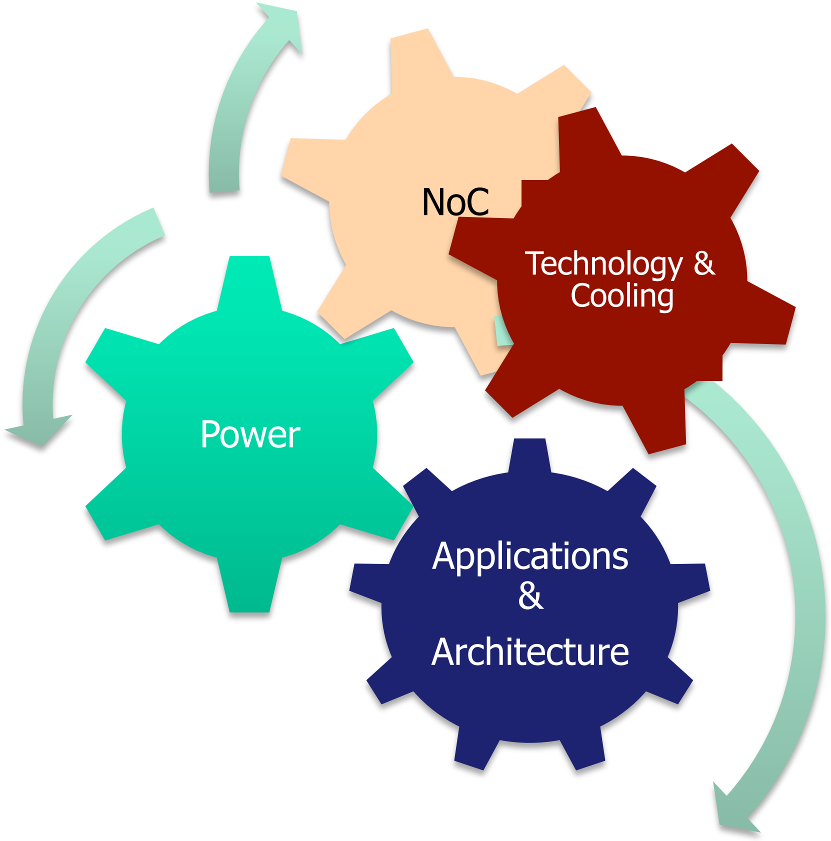
- Tradeoff between cooling capacity and 3D bandwidth
- Co-Design Exploration
  - Routing quality plays a key role

Parameters	Values
Chip Size	(1.65x1.65)cm^2
Number of CLBs	980K
Number of TSVs per 3D Switch Box	4
Horizontal Routing Channel BW	30
Micro-Channel Dimension ()	(50x100) μm^2
Fluid Velocity	1 m s^{-1}

Concluding Remarks

- We are going to see a reshaping of the boundaries between compute and memory
- System-level Co-design for memory-centric compute
- Exploration of network technologies: wireless, capacitive coupling, optics, etc.
- Expand the scope of traditional NoCs

Scaling Performance ⇨



Thank You

Questions?