
Exploring Emerging (Memory) Technologies in the HPC Co- Design Space



Jeffrey S. Vetter

Presented to
Clusters, Clouds, and Data for
Scientific Computing (CCDSC)
Lyon

4 Sep 2014



OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

**Georgia
Tech**



**College of
Computing**

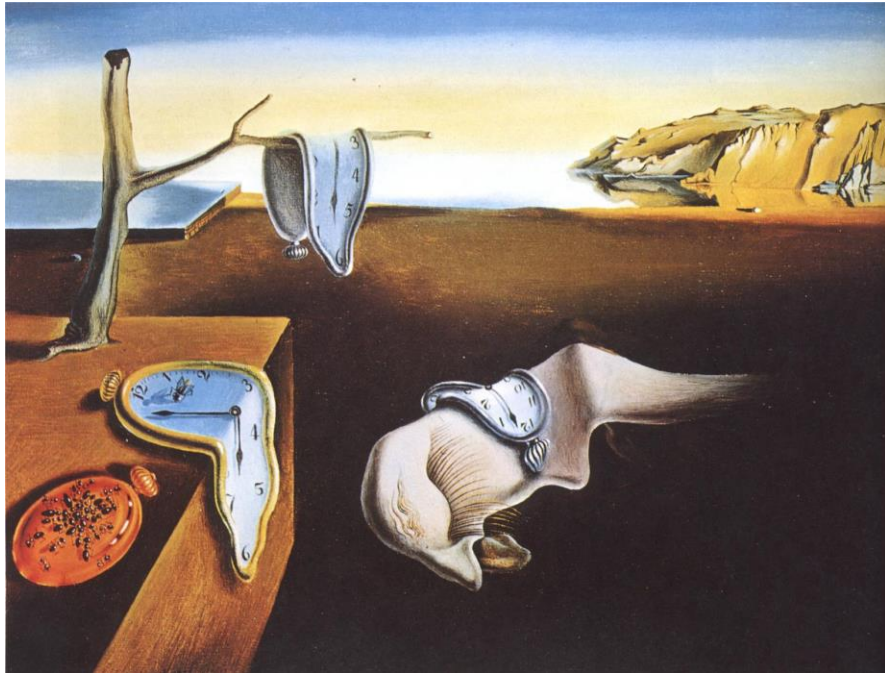
Computational Science and Engineering

<http://ft.ornl.gov> ♦ vetter@computer.org

Highlights

- New and Improved Memory systems are the next Big Thing
- Heterogeneous computing is here to stay
- Application characteristics (should) matter

New and Improved Memory Systems are the Next Big Thing

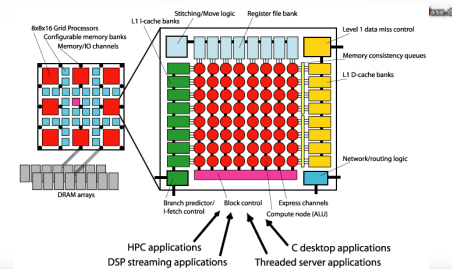
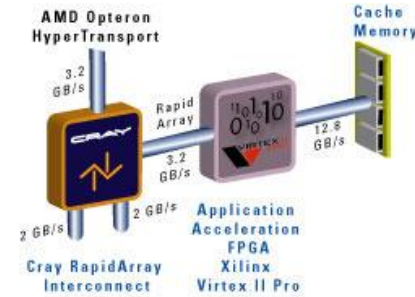
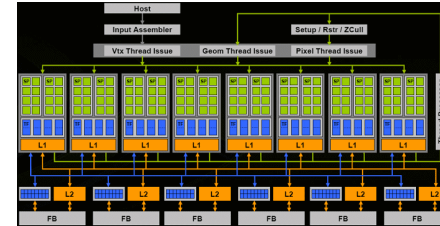
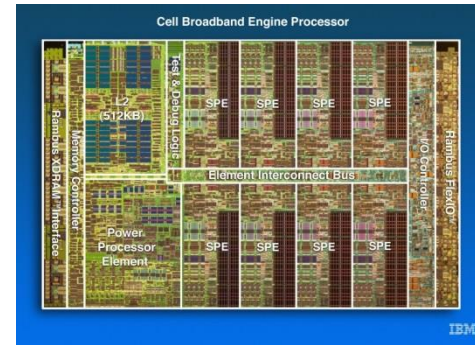


The Persistence of Memory

Earlier Experimental Computing Systems

- The past decade has started the trend away from traditional 'simple' architectures
- Mainly driven by facilities costs and successful (sometimes heroic) application examples
- Examples
 - Cell, GPUs, FPGAs, SoCs, etc
- Many open questions
 - Understand technology challenges
 - Evaluate and prepare applications
 - Recognize, prepare, enhance programming models

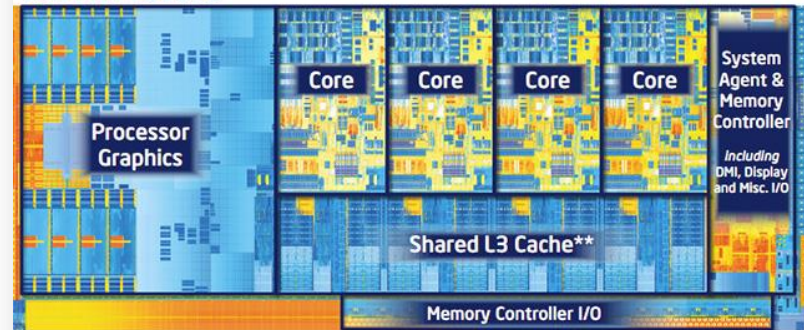
Popular architectures since ~2004



Emerging Computing Architectures – Future

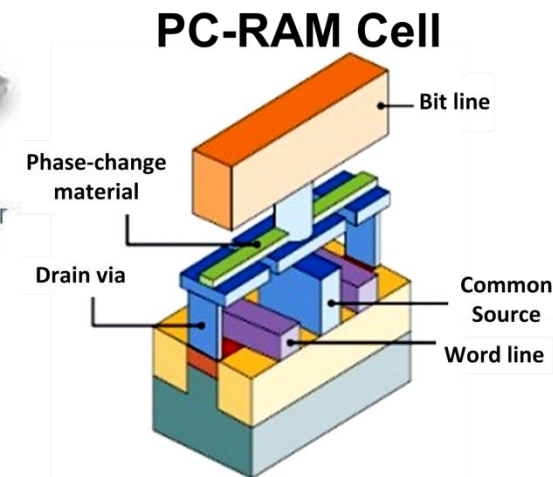
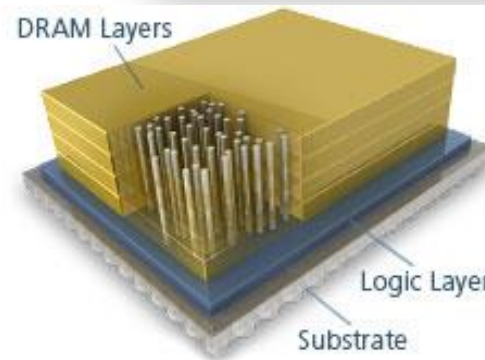
- Heterogeneous processing
 - Latency tolerant cores
 - Throughput cores
 - Special purpose hardware (e.g., AES, MPEG, RND)
 - Fused, configurable memory
- Memory
 - 2.5D and 3D Stacking
 - HMC, HBM, WIDEIO2, LPDDR4, etc
 - New devices (PCRAM, ReRAM)
- Interconnects
 - Collective offload
 - Scalable topologies
- Storage
 - Active storage
 - Non-traditional storage architectures (key-value stores)
- Improving performance and programmability in face of increasing complexity
 - Power, resilience

3rd Generation Intel® Core™ Processor:
22nm Process



New architecture with shared cache delivering more performance and energy efficiency

Quad Core die with Intel® HD Graphics 4000 shown above
Transistor count: 1.4Billion Die size: 160mm²
** Cache is shared across all 4 cores and processor graphics

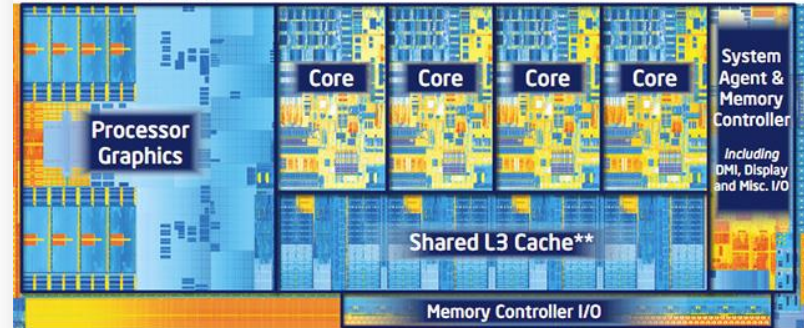


HPC (mobile, enterprise, embedded) computer design is more fluid now than in the past two decades.

Emerging Computing Architectures – Future

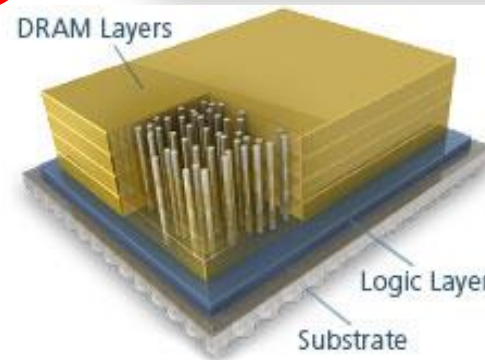
- Heterogeneous processing
 - Latency tolerant cores
 - Throughput cores
 - Special purpose hardware (e.g., AES, MPEG, RND)
 - Fused, configurable memory
- Memory
 - 2.5D and 3D Stacking
 - HMC, HBM, WIDEIO2, LPDDR4, etc
 - New devices (PCRAM, ReRAM)
- Interconnects
 - Collective offload
 - Scalable topologies
- Storage
 - Active storage
 - Non-traditional storage architectures (key-value stores)
- Improving performance and programmability in face of increasing complexity
 - Power, resilience

3rd Generation Intel® Core™ Processor:
22nm Process

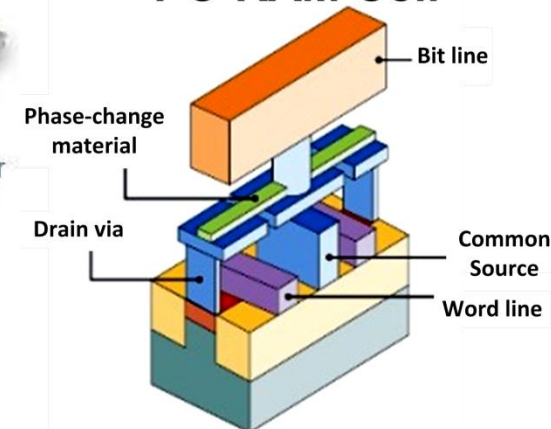


New architecture with shared cache delivering more performance and energy efficiency

Quad Core die with Intel® HD Graphics 4000 shown above
Transistor count: 1.4Billion Die size: 160mm²
** Cache is shared across all 4 cores and processor graphics



PC-RAM Cell



HPC (mobile, enterprise, embedded) computer design is more fluid now than in the past two decades.

Notional Exascale Architecture Targets

(From Exascale Arch Report 2009)

System attributes	2001	2010	“2015”		“2018”	
System peak	10 Tera	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	~0.8 MW	6 MW	15 MW		20 MW	
System memory	0.006 PB	0.3 PB	5 PB		32-64 PB	
Node performance	0.024 TF	0.125 TF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW		25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	16	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	416	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW		1.5 GB/s	150 GB/sec	1 TB/sec	250 GB/sec	2 TB/sec
MTTI		day	O(1 day)		O(1 day)	

Parallel I/O ??

NVRAM Technology Continues to Improve – Driven by Market Forces

designlines MEMORY

News & Analysis

3D NAND Production Starts at Samsung

Peter Clarke

8/6/2013 08:05 AM EDT

16 comments

Like 17 Tweet 7 Share 10 +1 3

LONDON — Samsung Electronics Co. Ltd. has begun production of its 3D NAND memory technology, which is expected to be a significant improvement over the current 2D NAND technology.

The memory conversion in the vertical reliability conversion in a press

The technology did not convert whether in 2D memory

The conversion improves is suitable for applications drives.

designlines MEMORY

News & Analysis

3D NAND Transition: 15nm Process Technology Takes Shape

Gary Hilson

5/13/2014 08:15 AM EDT

5 comments

Like 15 Tweet 6 Share 6

TORONTO — With 3D NAND unlikely to make it at least 2015, SanDisk and its flash foundry recently announced 15nm process technology for flash.

SanDisk's 17-nm technology will be applied to

Original URL: http://www.theregister.co.uk/2013/11/01/hp_memristor_2018/

HP 100TB Memristor drives by 2018 – if you're lucky, admits tech titan

Universal memory slow in coming

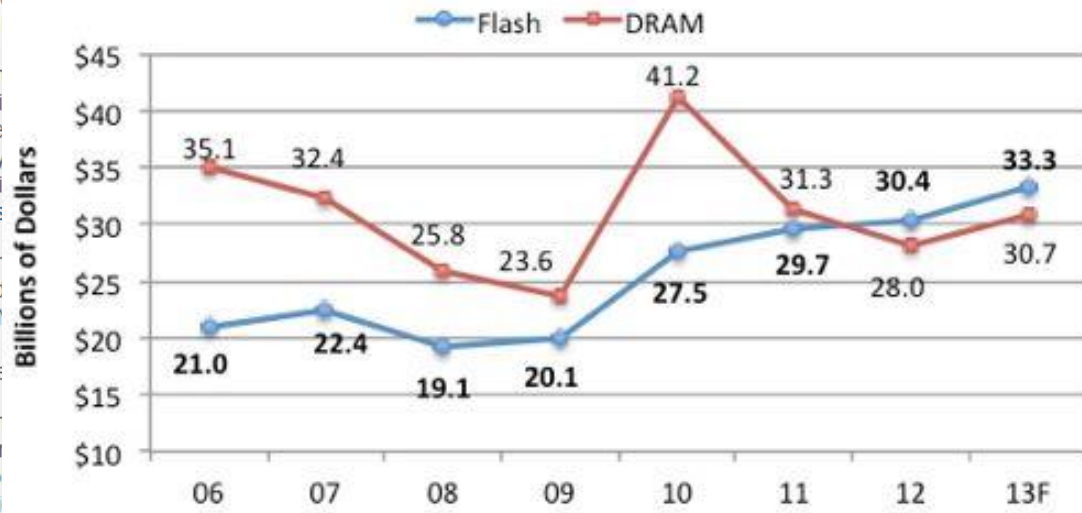
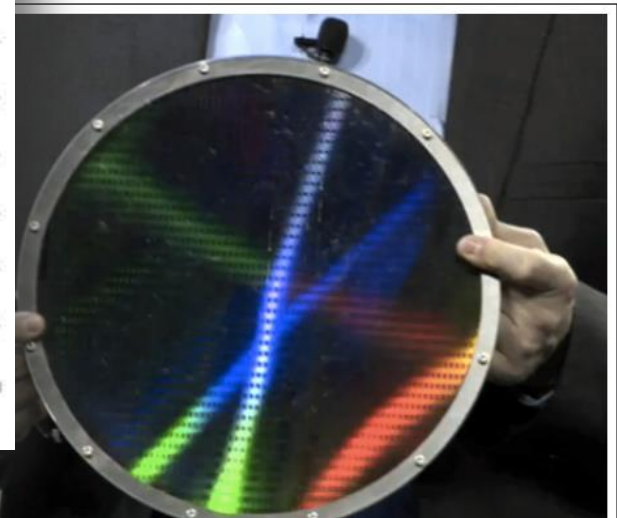
By Chris Mellor

Posted in Storage, 1st November 2013 02:28 GMT

Blocks and Files HP has warned *EI Reg* not to get its hopes up too high after the tech titan's CTO Martin Fink suggested StoreServ arrays could be packed with 100TB Memristor drives come 2018.

In five years, according to Fink, DRAM and NAND scaling will hit a wall, limiting the maximum capacity of the technologies: process shrinks will come to a shuddering halt when the memories' reliability drops off a cliff as a side effect of reducing the size of electronics on the silicon dies.

The HP answer to this scaling wall is Memristor, its flavour of resistive RAM technology that is supposed to offer faster speed and better-than-NAND storage density. Fink claimed at an HP Discover event that Memristor devices will be ready by the time flash NAND hits its limit in five years. He also showed a Memristor wafer, adding that it could have a 1.5PB capacity by the end of the decade.



http://www.eetasia.com/STATIC/ARTICLE_IMAGES/201212/EEOL_2012DEC28_STORAGE_ROLE_FOR_NAND_01.jpg

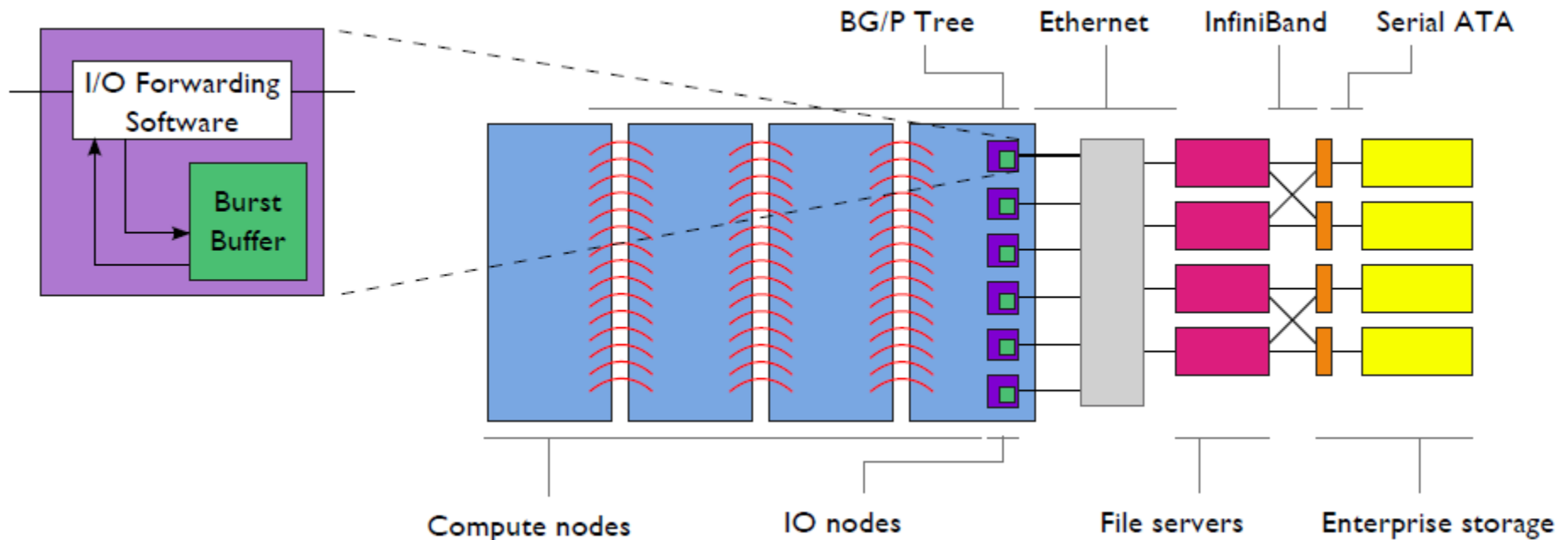
The V-NAND component is a 3D NAND technology that is expected to be a significant improvement over the current 2D NAND technology.

Blackcomb: Comparison of emerging memory technologies

	SRAM	DRAM	eDRAM	2D NAND Flash	3D NAND Flash	PCRAM	STTRAM	2D ReRAM	3D ReRAM
Data Retention	N	N	N	Y	Y	Y	Y	Y	Y
Cell Size (F ²)	50-200	4-6	19-26	2-5	<1	4-10	8-40	4	<1
Minimum F demonstrated (nm)	14	25	22	16	64	20	28	27	24
Read Time (ns)	< 1	30	5	10 ⁴	10 ⁴	10-50	3-10	10-50	10-50
Write Time (ns)	< 1	50	5	10 ⁵	10 ⁵	100-300	3-10	10-50	10-50
Number of Rewrites	10 ¹⁶	10 ¹⁶	10 ¹⁶	10 ⁴ -10 ⁵	10 ⁴ -10 ⁵	10 ⁸ -10 ¹⁰	10 ¹⁵	10 ⁸ -10 ¹²	10 ⁸ -10 ¹²
Read Power	Low	Low	Low	High	High	Low	Medium	Medium	Medium
Write Power	Low	Low	Low	High	High	High	Medium	Medium	Medium
Power (other than R/W)	Leakage	Refresh	Refresh	None	None	None	None	Sneak	Sneak
Maturity									

Q: How do we integrate this new memory into a system and how do we expose it to applications?

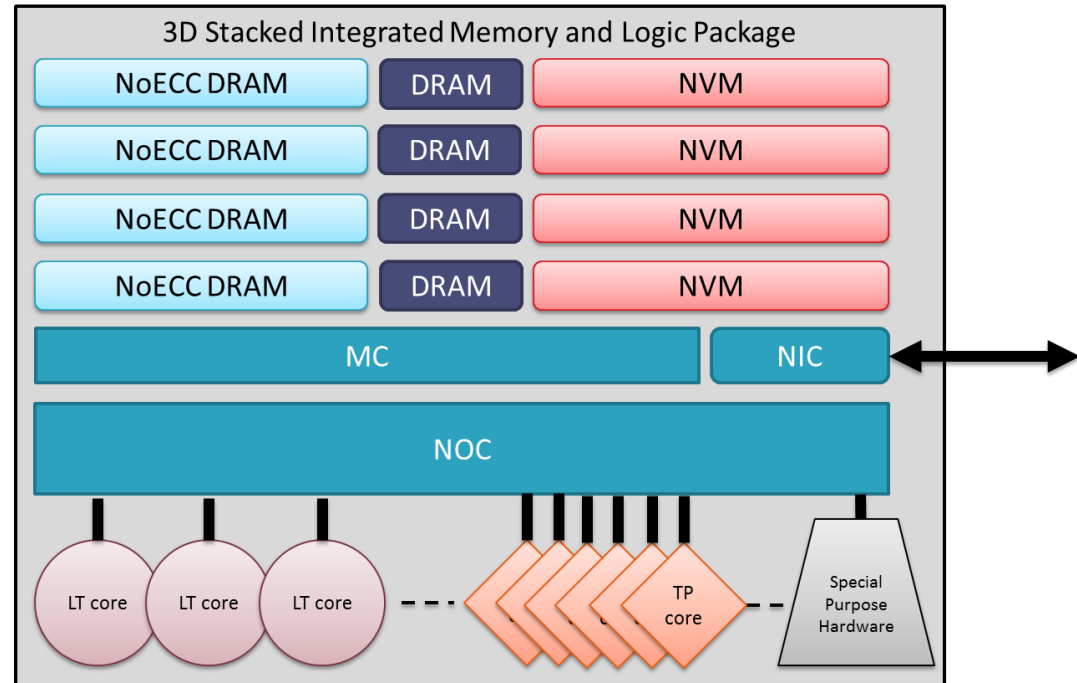
Early Uses of NVRAM: Burst Buffers



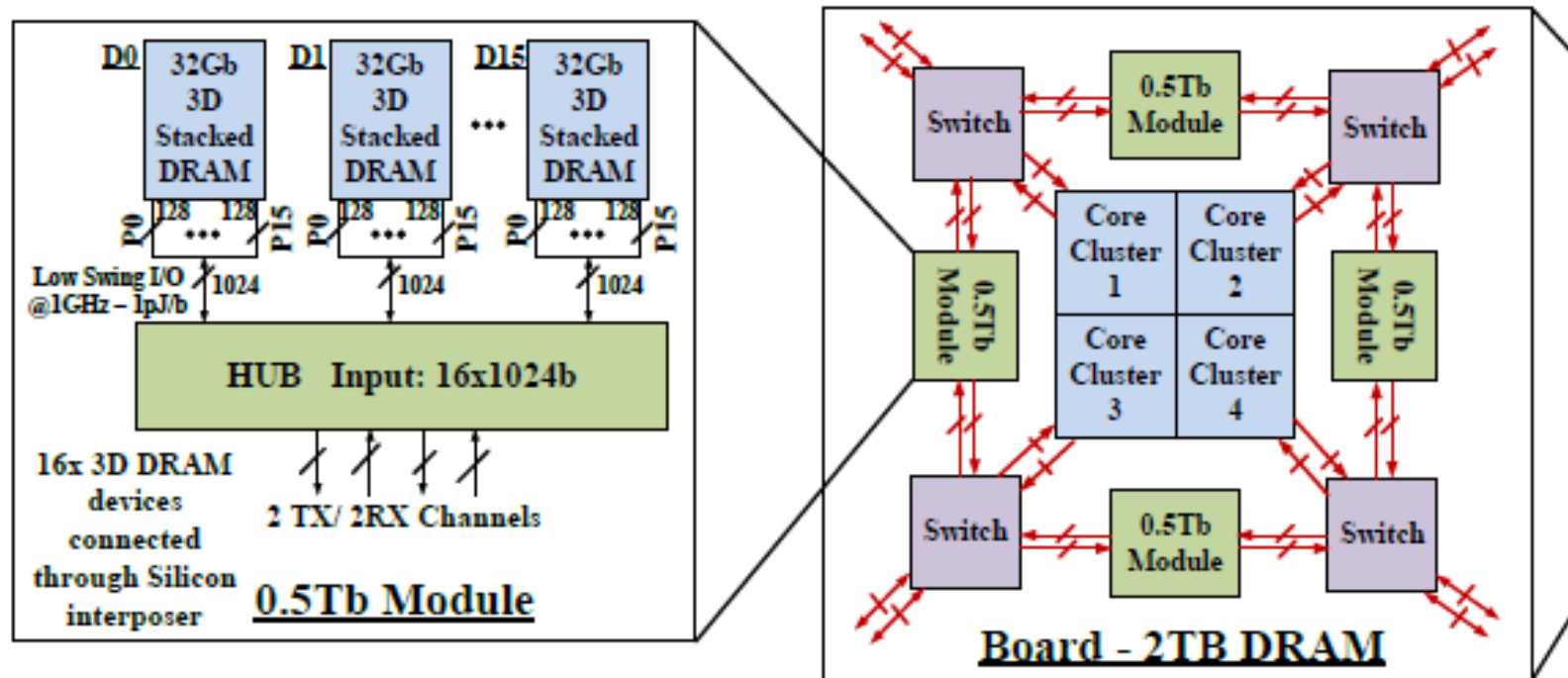
N. Liu, J. Cope, P. Carns, C. Carothers, R. Ross, G. Grider, A. Crume, and C. Maltzahn, "On the role of burst buffers in leadership-class storage systems," Proc. IEEE 28th Symposium on Mass Storage Systems and Technologies (MSST), 2012, pp. 1-11,

Notional Future Node Architecture

- Stacking increases local bandwidth, reduces power costs
 - Very high bandwidth, low latency
- NVM to increase memory capacity
- Mix of cores to provide different capabilities
- Integrated network interface



Tradeoffs in Exascale Memory Architectures

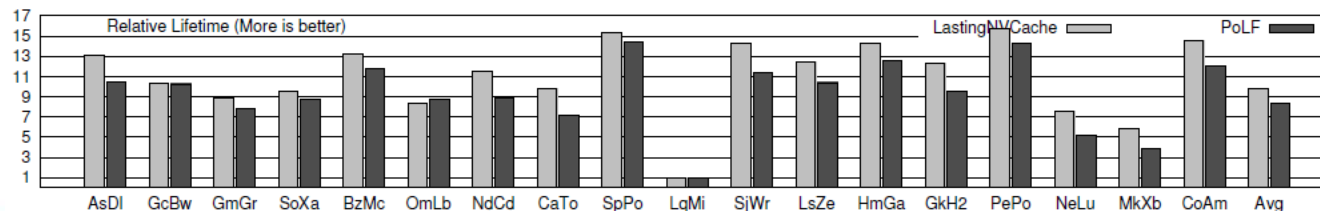


- Understanding the tradeoffs
 - ECC type, row buffers, DRAM physical page size, bitline length, etc

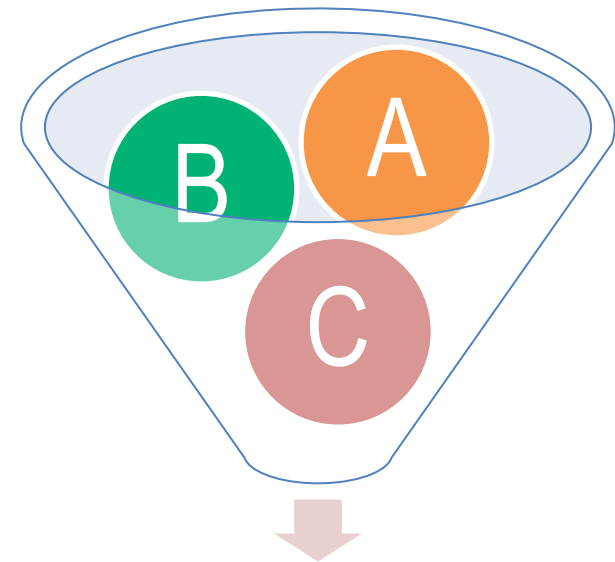
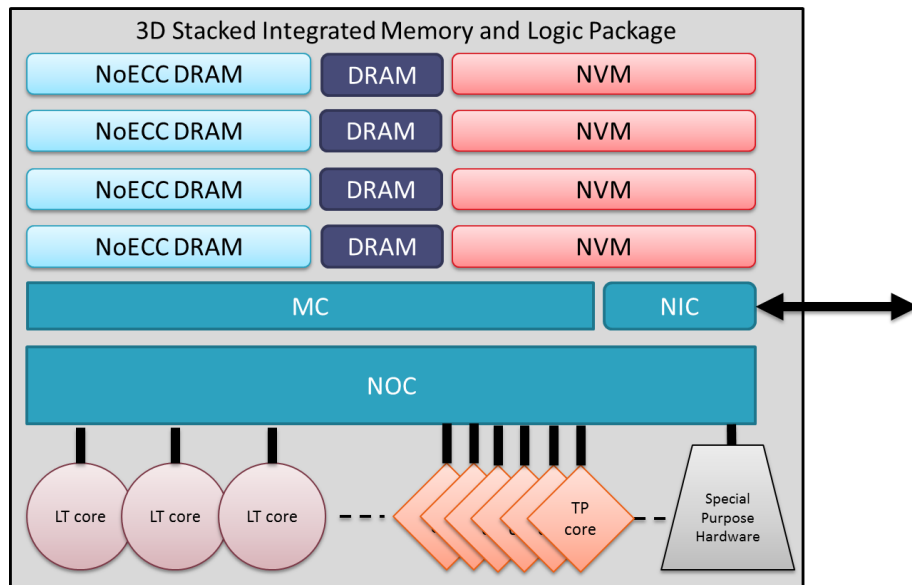
Increasing Lifetime of NVM Caches

- Problem
 - Write endurance of NVM caches (e.g. Resistive RAM) is very small and cache management policies introduce large write-variation => cache lifetime becomes small
- Solution
 - We propose LastingNVCache, a technique to improve NVM lifetime
 - After a fixed number of writes on a block, it is flushed. Thus, future writes are redirected from a hot-block to a cold-block
 - This leads to intra-set wear-leveling which improves the cache lifetime
- Recent results
 - On Sniper simulator with 1, 2 and 4-core configuration, SPEC06 and DoE workloads
 - LastingNVCache significantly improves lifetime and outperforms a recent prior work (PoLF)
 - It incurs smaller performance and energy overhead than PoLF
- Impact
 - An important step towards making NVMs practical and universal memory solution
 - Our technique can also be used for mitigating NVM cache write attacks.

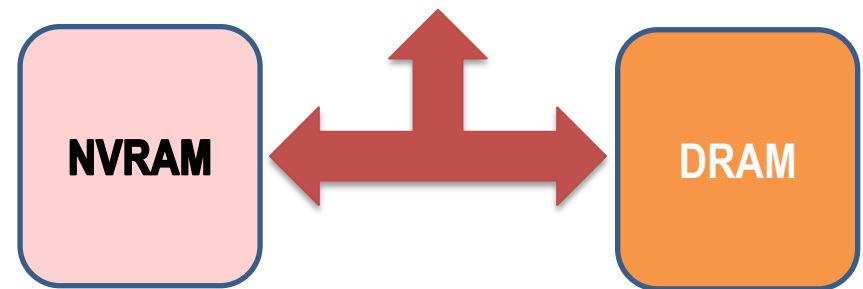
Accepted in ISVLSI 2014 “LastingNVCache: A Technique for Improving the Lifetime of Non-volatile”



New hybrid memory architectures: What is the ideal organizations for our applications?



Natural separation of applications objects?



Observations: Numerous characteristics of applications are a good match for byte-addressable NVRAM

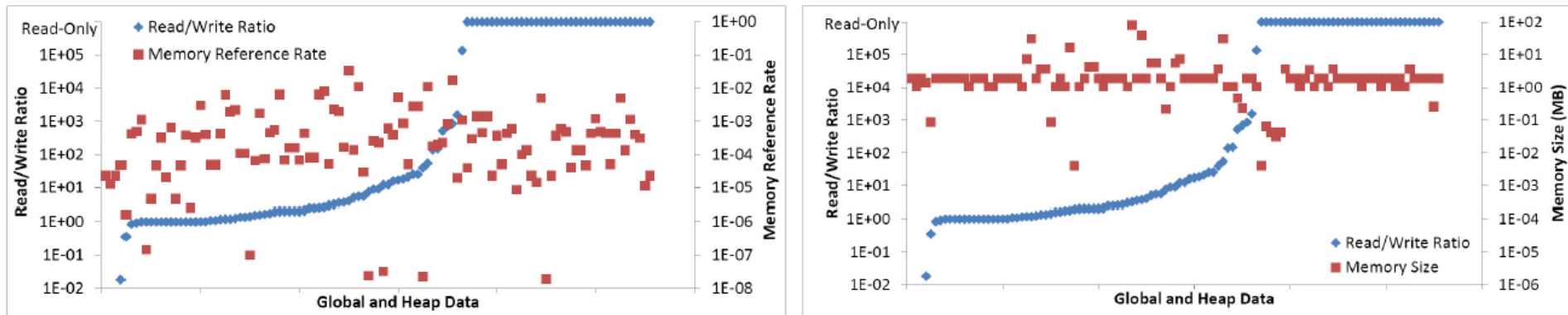


Figure 3: Read/write ratios, memory reference rates and memory object sizes for memory objects in Nek5000

- Many lookup, index, and permutation tables
- Inverted and 'element-lagged' mass matrices
- Geometry arrays for grids
- Thermal conductivity for soils
- Strain and conductivity rates
- Boundary condition data
- Constants for transforms, interpolation
- ...

Programming Interfaces Example: NV-HEAPS

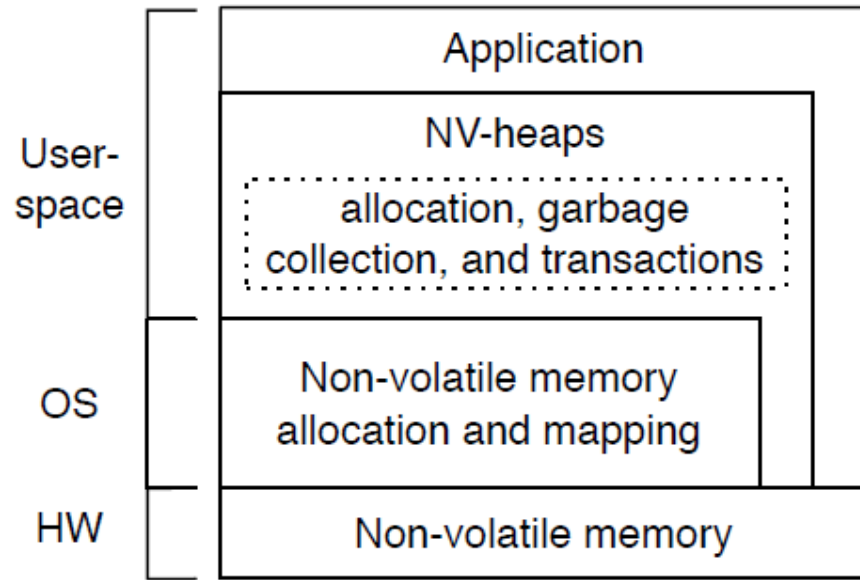


Figure 1. The NV-heap system stack This organization allows read and write operations to bypass the operating system entirely.

```
class NVList : public NVObject {
    DECLARE_POINTER_TYPES(NVList);
public:
    DECLARE_MEMBER(int, value);
    DECLARE_PTR_MEMBER(NVList::NVPtr, next);
};

void remove(int k)
{
    NVHeap * nv = NVHOpen("foo.nvheap");
    NVList::VPtr a =
        nv->GetRoot<NVList::NVPtr>();
    AtomicBegin {
        while(a->get_next() != NULL) {
            if (a->get_next()->get_value() == k) {
                a->set_next(a->get_next()->get_next());
            }
            a = a->get_next();
        }
    } AtomicEnd;
}
```

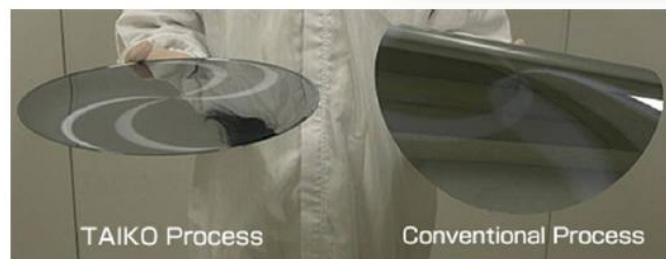
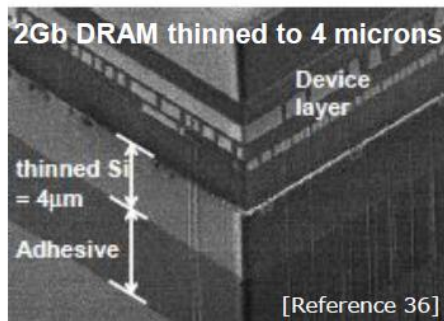
Figure 2. NV-heap example A simple NV-heap function that atomically removes all links with value k from a non-volatile linked list.

In other news: Stacking Technologies Continue to Improve

- HMC, HBM, etc

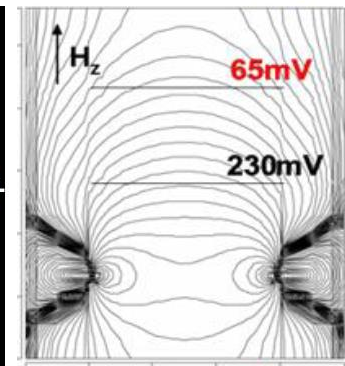
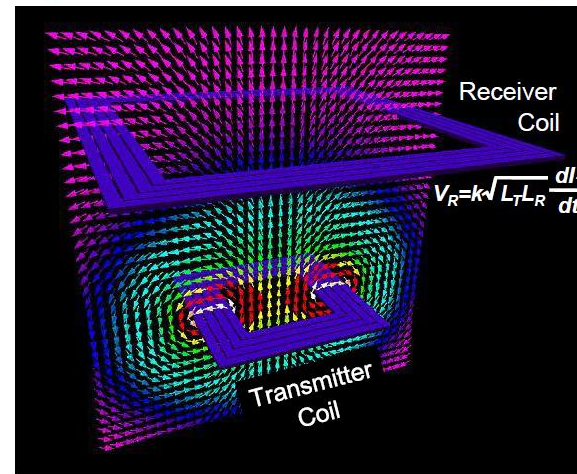
Ultra-Thin 4 μ wafer breakthrough

- Wafer thinning has been stuck at ~40 μ due to “Gettering”
 - Barrier was due in part to loss of the “gettering effect” at smaller diameter grinding, causing impurities affecting device performance (partially)
- DISCO Corporation solution can now thin to a few microns
 - DISCO introduced a “Gettering Dry Polish” wheel which forms gettering pits allowing thinning of wafer silicon to a few microns without device degradation
- Example: DRAM silicon thinned to 4 microns
 - See “Ultra Thinning down to 4 μ m using 300-mm Wafer proven by 4 3D Multi-stack WOW Applications.”[36] They concluded “No degradation in terms of retention characteristics and distribution employing 2 Gb DRAM wafer was found after ultra-thinning.”



Ultra-thin wafers can be handled (from DISCO website)

Communication is via magnetic field



$$\mu_{Si} = \mu_{SiO_2} = 1$$

Can easily induce a 200 mV signal in receiver coil.

Magnetic field can pass through silicon, including over active circuitry.

August 11, 2014

Hot Chips 26 – ThruChip Wireless Connections

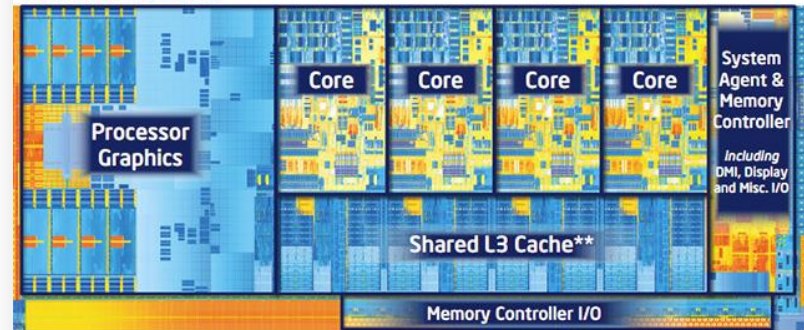
13

Heterogeneous computing is
here to stay

Emerging Computing Architectures – Future

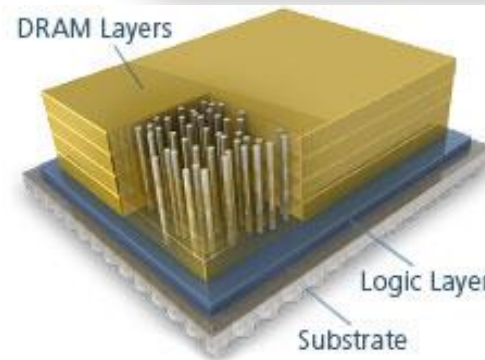
- Heterogeneous processing
 - Latency tolerant cores
 - Throughput cores
 - Special purpose hardware (e.g., AES, MPEG, RND)
 - Fused, configurable memory
- Memory
 - 2.5D and 3D Stacking
 - HMC, HBM, WIDEIO2, LPDDR4, etc
 - New devices (PCRAM, ReRAM)
- Interconnects
 - Collective offload
 - Scalable topologies
- Storage
 - Active storage
 - Non-traditional storage architectures (key-value stores)
- Improving performance and programmability in face of increasing complexity
 - Power, resilience

3rd Generation Intel® Core™ Processor:
22nm Process

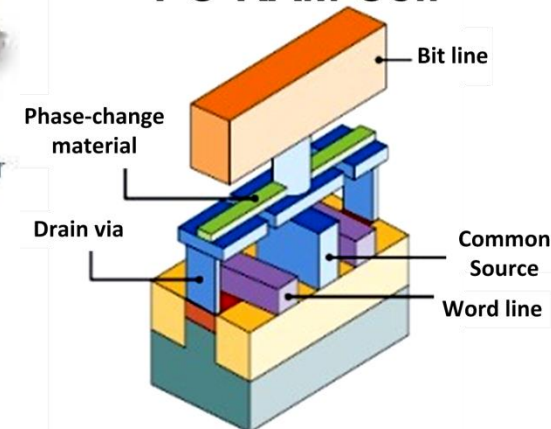


New architecture with shared cache delivering more performance and energy efficiency

Quad Core die with Intel® HD Graphics 4000 shown above
Transistor count: 1.4Billion Die size: 160mm²
** Cache is shared across all 4 cores and processor graphics



PC-RAM Cell



HPC (mobile, enterprise, embedded) computer design is more fluid now than in the past two decades.

Contemporary HPC Architectures

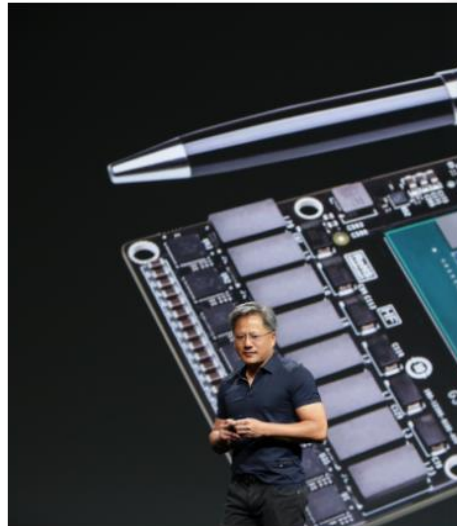
Date	System	Location	Comp	Comm	Peak (PF)	Power (MW)
2009	Jaguar; Cray XT5	ORNL	AMD 6c	Seastar2	2.3	7.0
2010	Tianhe-1A	NSC Tianjin	Intel + NVIDIA	Proprietary	4.7	4.0
2010	Nebulae	NSCS Shenzhen	Intel + NVIDIA	IB	2.9	2.6
2010	Tsubame 2	TiTech	Intel + NVIDIA	IB	2.4	1.4
2011	K Computer	RIKEN/Kobe	SPARC64 VIIIfx	Tofu	10.5	12.7
2012	Titan; Cray XK6	ORNL	AMD + NVIDIA	Gemini	27	9
2012	Mira; BlueGeneQ	ANL	SoC	Proprietary	10	3.9
2012	Sequoia; BlueGeneQ	LLNL	SoC	Proprietary	20	7.9
2012	Blue Waters; Cray	NCSA/UIUC	AMD + (partial) NVIDIA	Gemini	11.6	
2013	Stampede	TACC	Intel + MIC	IB	9.5	5
2013	Tianhe-2	NSCC-GZ (Guangzhou)	Intel + MIC	Proprietary	54	~20

Recent announcements (1)

Nvidia and IBM create GPU interconnect for faster supercomputing

"NVLink" shares up to 80GB of data per second between CPUs and GPUs.

by Jon Brodtkin - Mar 25 2014, 2:45pm EST



Nvidia CEO Jen-Hsun Huang introduces Pascal at today's GPU Tech

Nvidia

Nvidia and IBM have developed an interconnect that will be units, letting GPUs and CPUs share data five times faster than today. The fatter pipe will let data flow between the CPU and GPU compared to 16GB per second today.

It Begins: AMD Announces Its First ARM Based Server SoC, 64-bit/8-core Opteron A1100

by Anand Lal Shimpi on January 28, 2014 6:35 PM EST

Posted in CPUs IT Computing Enterprise enterprise CPUs AMD Opteron Opteron A1100 ARM

123
Comments

+ Add A
Comment

"SEATTLE" 64-BIT ARM SERVER PROCESSOR FIRST 28NM ARM SERVER CPU TO SAMPLE IN MARCH

AMD

- Industry's only 64-bit ARM Server SoC from a proven server processor supplier
 - The most server experience of any ARM licensee
 - Server class IP blocks—no other competitor has
- CPU code named "Seattle"
 - 2-4x the performance of AMD Opteron™ X-Series with significant improvement in compute per watt³
 - 8 core SoCs with 128 GB DRAM support
 - Based on ARM Cortex™-A57 cores at > 2 GHz
 - Extensive offload engines for better power efficiency and reduced CPU loading
 - Server caliber encryption and compression
 - Legacy Networking: Integrated 10GbE
 - Storage: High port-count storage interfaces optimized for big data

SAMPLING IN A FEW WEEKS

12 | CHANGING INFRASTRUCTURE LANDSCAPE | JANUARY 2014 | CONFIDENTIAL

Around 15 months ago, AMD announced that it would be building a 64-bit ARM server processor in 2014. Less than a month into 2014, AMD made good on its promise with the A1100: a 64-bit ARM Cortex A57 based SoC.

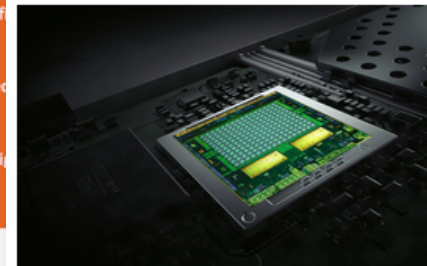
The Opteron A1100 features either 4 or 8 AMD Cortex A57 cores. AMD is not talking about harvested die to make up the quad-core configuration. They are going away entirely, but since we're at very early stages of talking about the product, bets are going on. Each core will run at a frequency somewhere in the ballpark of 2 GHz. The process at Global Foundries.

Nvidia Jetson TK1 mini supercomputer is up for pre-order

Will ship on 15 May

By Lee Bell

Fri May 02 2014, 11:38



NVIDIA'S JETSON TK1 mini supercomputer development kit is now up for pre-order, priced at \$192.

Despite Nvidia having announced on its blog that it is "now shipping", the development kit that is powered by a Tegra K1 chip won't actually ship until 15 May.

Claiming to be "the world's first mobile supercomputer", the Jetson TK1 kit is built for embedded systems to aid the development of self-driving cars.

computers attempting to simulate human recognition of physical objects, such as robots and self-driving cars.

Speaking at the GPU Technology Conference (GTC) in March, Nvidia co-founder and CEO Jen Hsun Huang described it as capable of running anything the GeForce GTX Titan Z graphics card can run, but at a slower pace.

With a total performance of 326 GFLOPS, the Jetson TK1 should be more powerful than the Raspberry Pi board, which delivers just 24 GFLOPS, but will retail for much more, costing \$192 in the US - a number that matches the number of cores in the Tegra K1 processor that Nvidia launched at CES in Las Vegas in January.

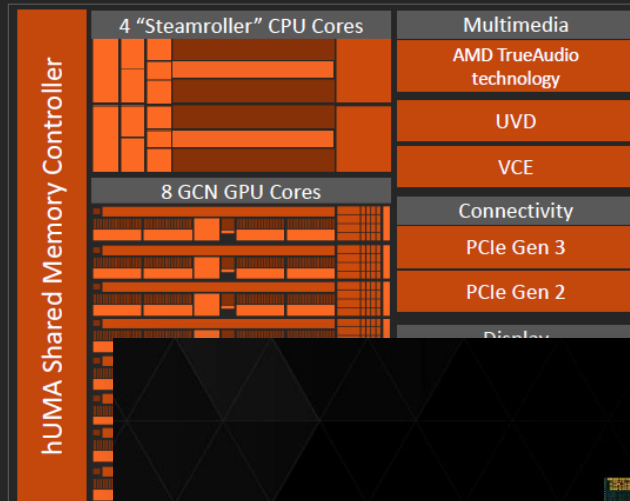
"The Jetson TK1 also comes with this new SDK called Vision Works. Stacked onto CUDA, it comes with a whole bunch of primitives whether it's recognising corners or detecting edges, or it could be classifying objects.

Parameters are loaded into this Vision Works primitives system and all of a sudden it recognises objects," Huang said on stage during the Jetson TK1 launch.

Recent announcements (2)

A-SERIES REDEFINES COMPUTE

Kaveri

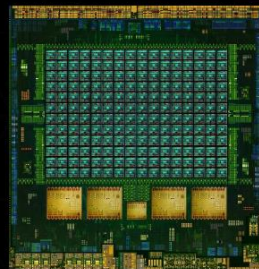


TEGRA K1

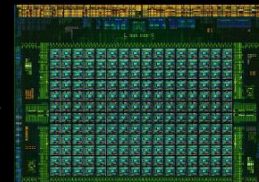
192-core
Kepler-Class Chip

3 | Applying AM

One Chip – Two Versions



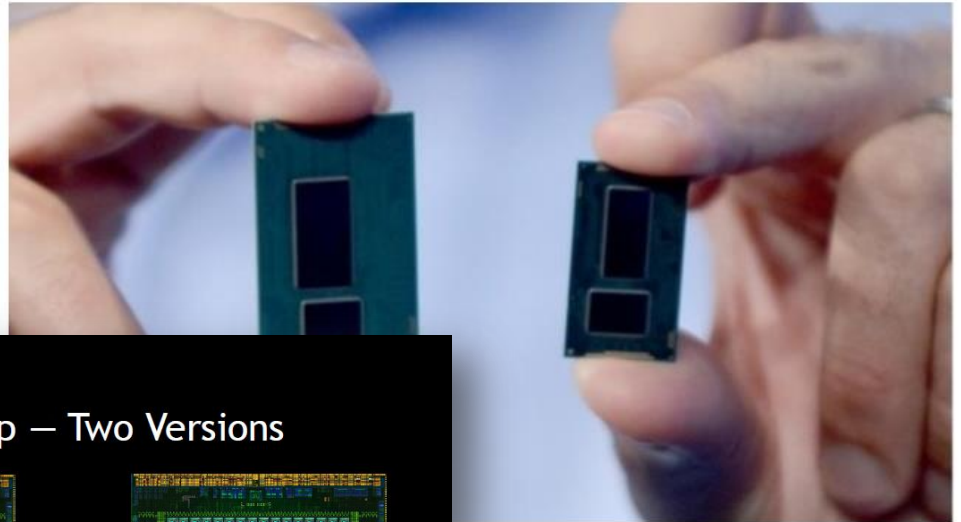
Pin
Compatible



Quad A15 CPUs
32-bit
3-way Superscalar
Up to 2.3GHz
32K+32K L1\$

Intel's 14nm Broadwell GPU takes shape, indicates major improvements over Haswell

By Sebastian Anthony on November 5, 2013 at 10:21 am | [16 Comments](#)



Ahead of its 2014 launch, Intel has started open-sourcing the Linux driver for Broadwell's GPU. Broadwell is the 14nm die shrink of Intel's

Intel Mates FPGA With Future Xeon Server Chip

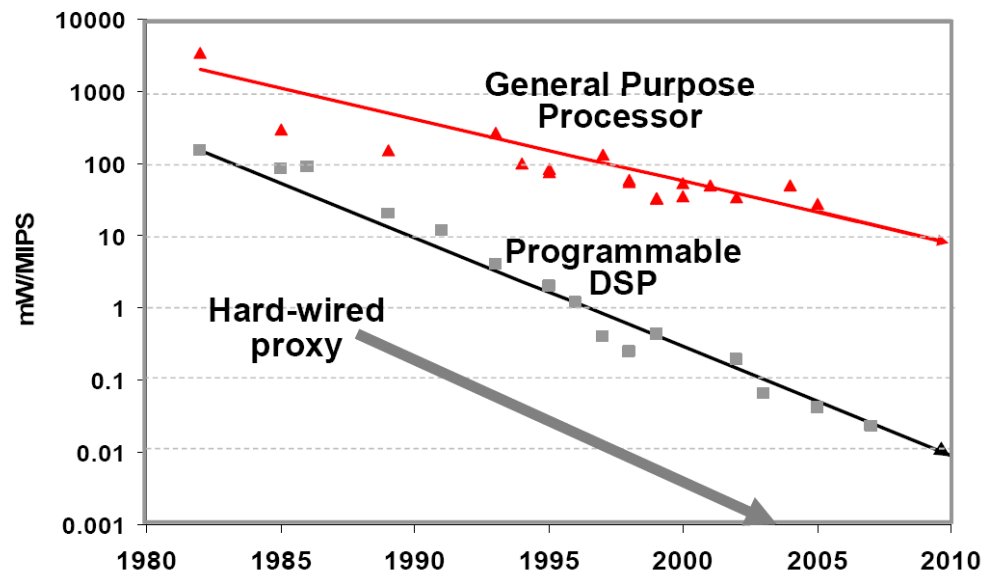
June 18, 2014 by Timothy Prickett Morgan



Intel is taking field programmable gate arrays seriously as a means of accelerating applications and has crafted a hybrid chip that marries an FPGA to a Xeon E5 processor and puts them in the same processor socket.

Mobile/Embedded Designers have Traveled this Path

No single architecture solves all power problems



- Industry has debated merits of each architecture for decades...
- Combination of all approaches optimizes power and performance

AMD Llano: Eliminating PCIe will Change Relevant Apps

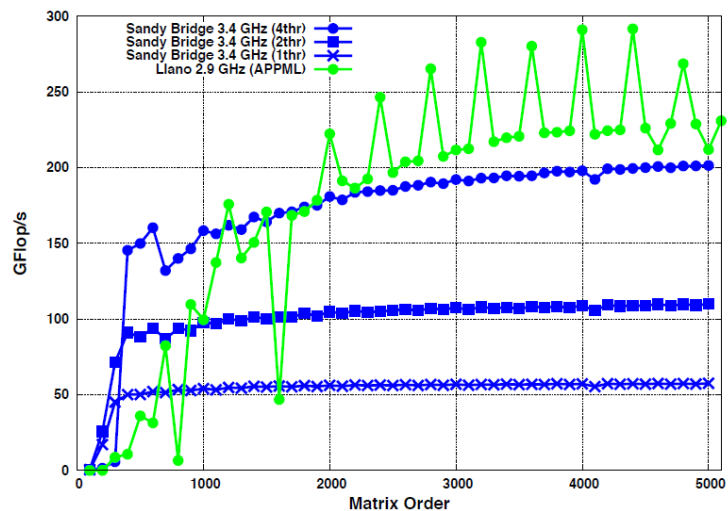
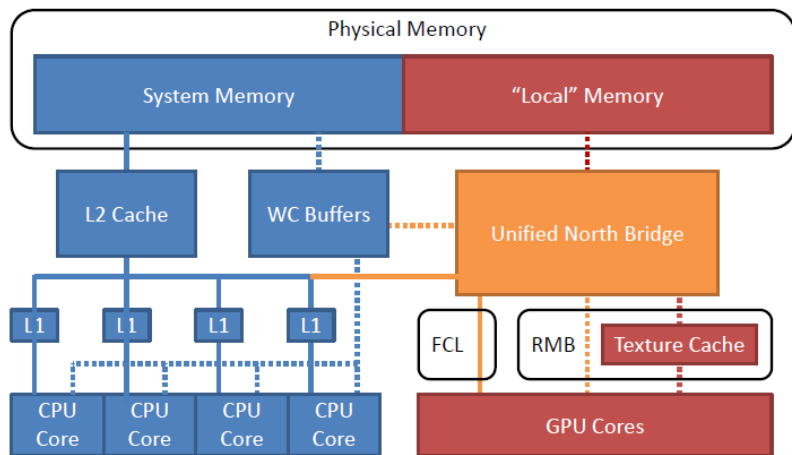
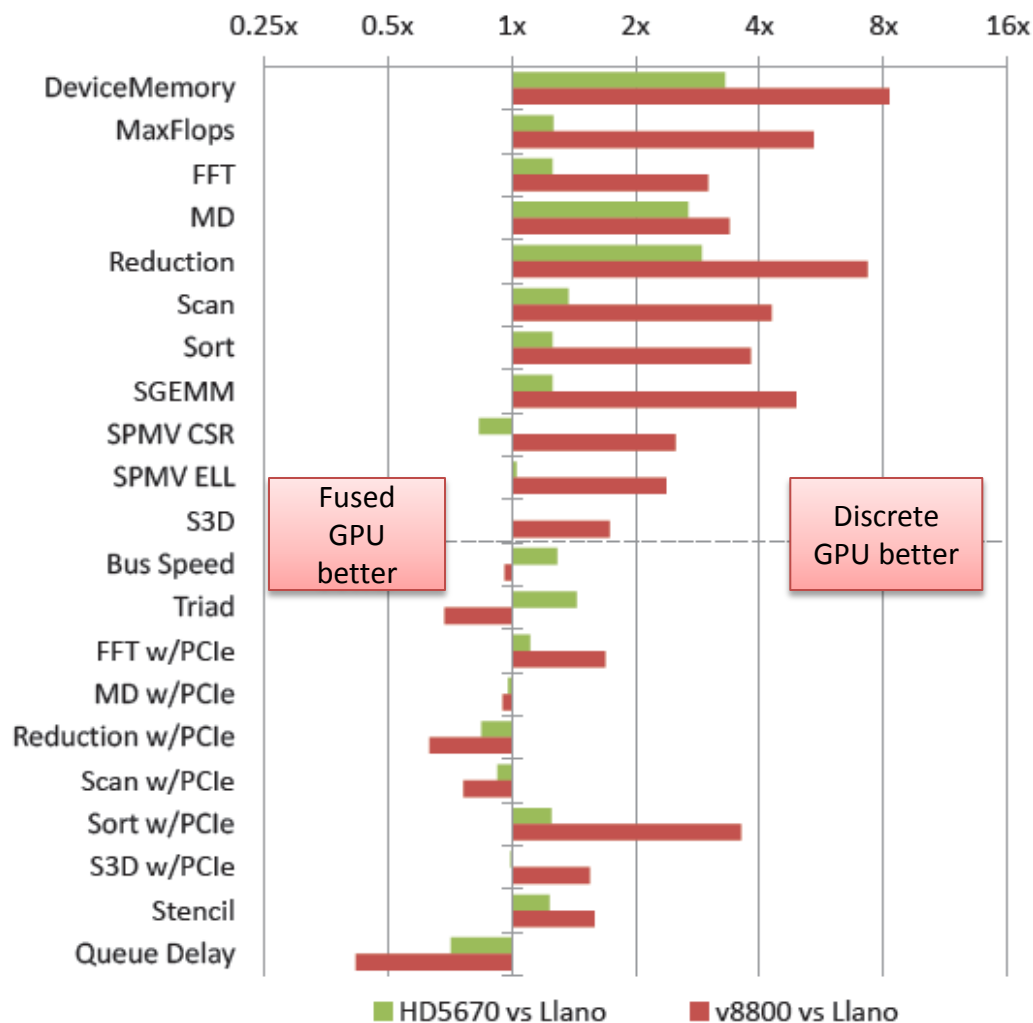
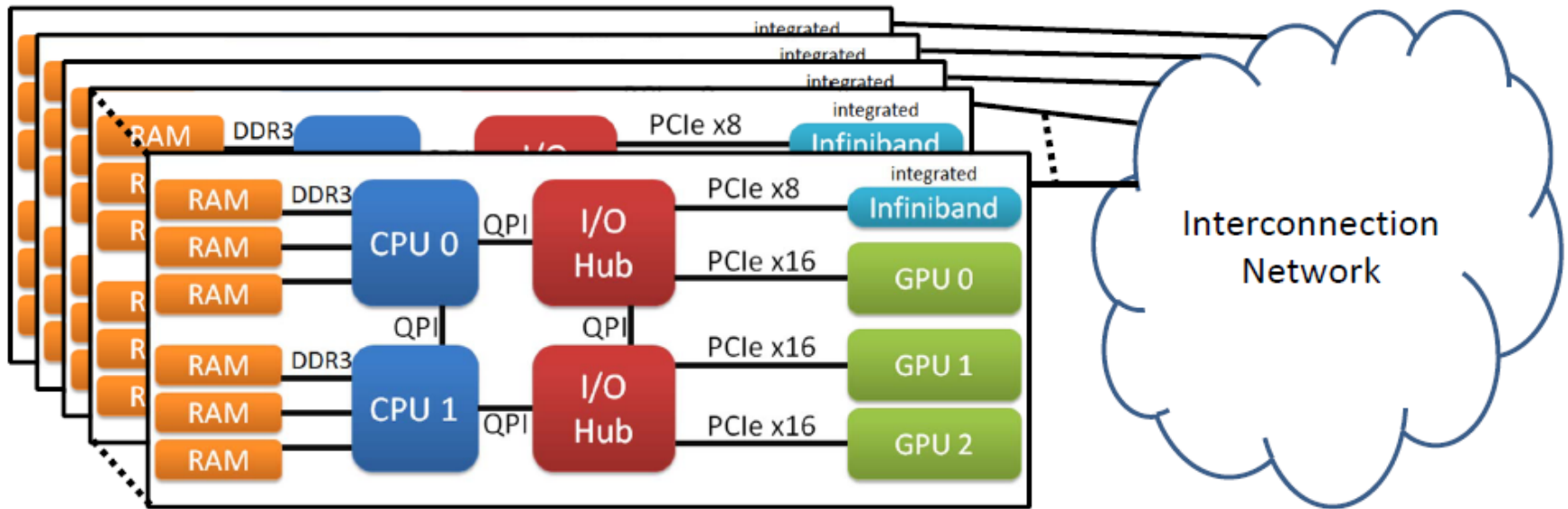


Figure 3: SGEMM Performance (one, two, and four CPU threads for Sandy Bridge and the OpenCL-based AMD APPML for Llano's fGPU)



K. Spafford, J.S. Meredith, S. Lee, D. Li, P.C. Roth, and J.S. Vetter, "The Tradeoffs of Fused Memory Hierarchies in Heterogeneous Architectures," in ACM Computing Frontiers (CF). Cagliari, Italy: ACM, 2012. Note: Both SB and Llano are consumer, not server, parts.

Applications must use a mix of programming models for these architectures



MPI

Low overhead

Resource contention

Locality

OpenMP, Pthreads

SIMD

NUMA

OpenACC, CUDA, OpenCL, OpenMP4, ...

Memory use,
coalescing

Data orchestration

Fine grained
parallelism

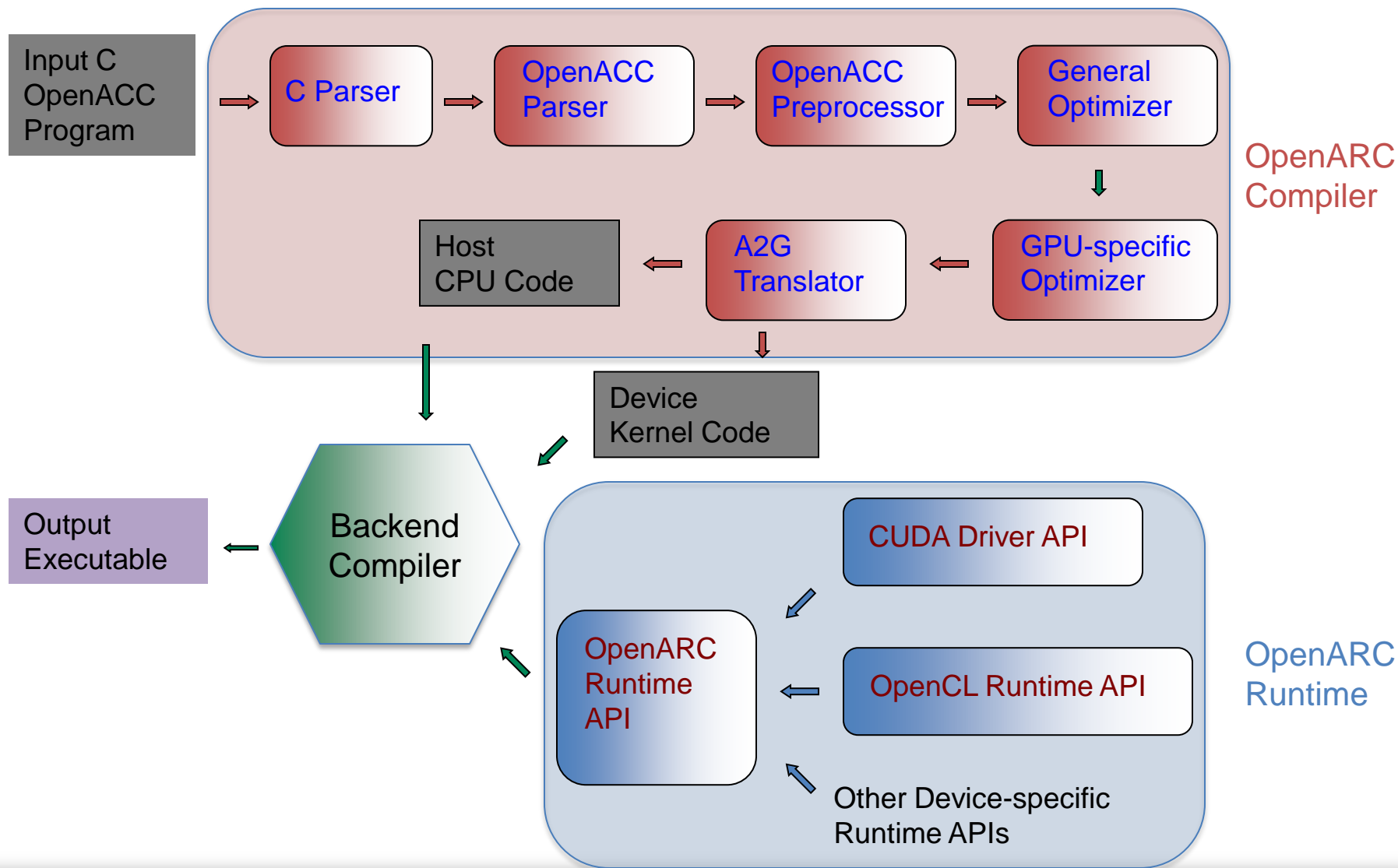
Hardware features

Realizing performance portability across contemporary heterogeneous architectures

Table 1: Comparison of Heterogeneous Architectures

Property	CUDA	GCN	MIC
Programming models	CUDA, OpenCL	OpenCL, C++ AMP	OpenCL, Cilk, TBB, LEO OpenMP
Thread Scheduling	Hardware	Hardware	Software
User Managed Cache	Yes	Yes	No
Global Synchronization	No	No	Yes
L2 Cache Type	Shared	Private per core	Private per core
L2 Total Size	upto 1.5MB	upto 0.5MB	25MB
L2 Line-size	128	64	64
L1 Data Cache	Read-only + Read-write	Read-only	Read-write
Native Mode	No	No	Yes

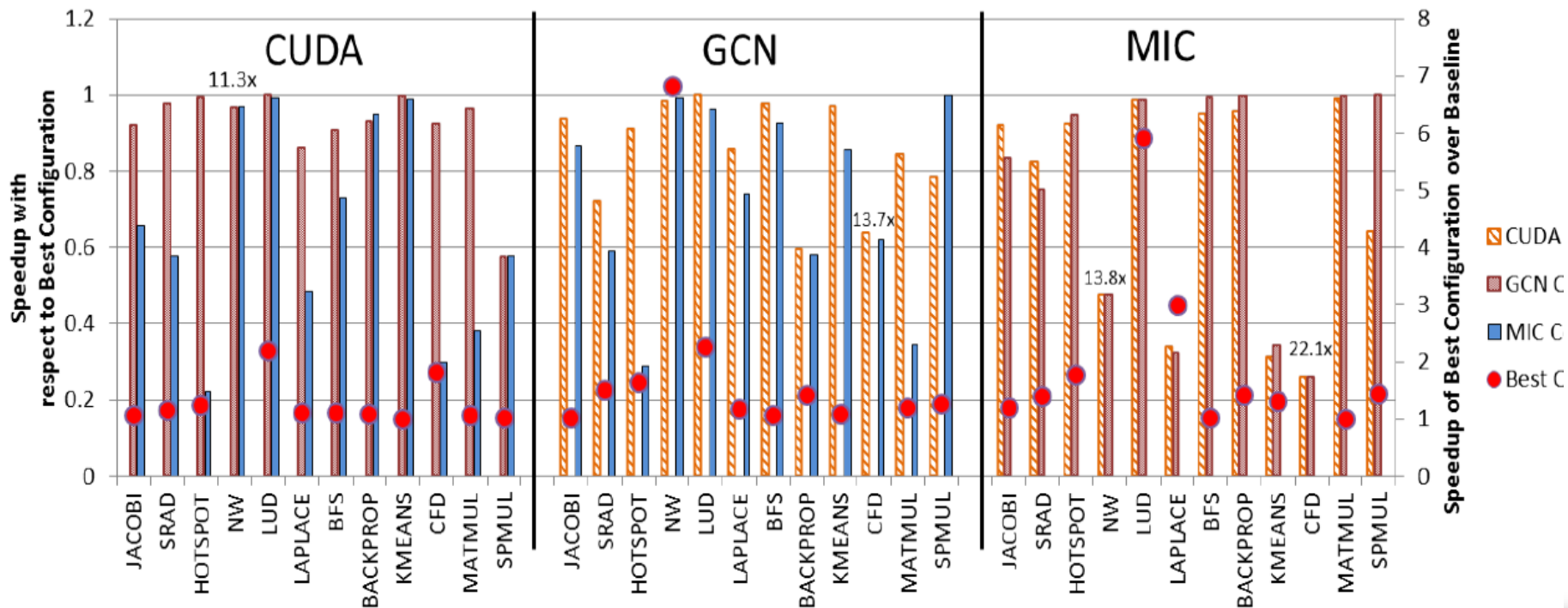
OpenARC System Architecture



Performance Portability is critical and challenging

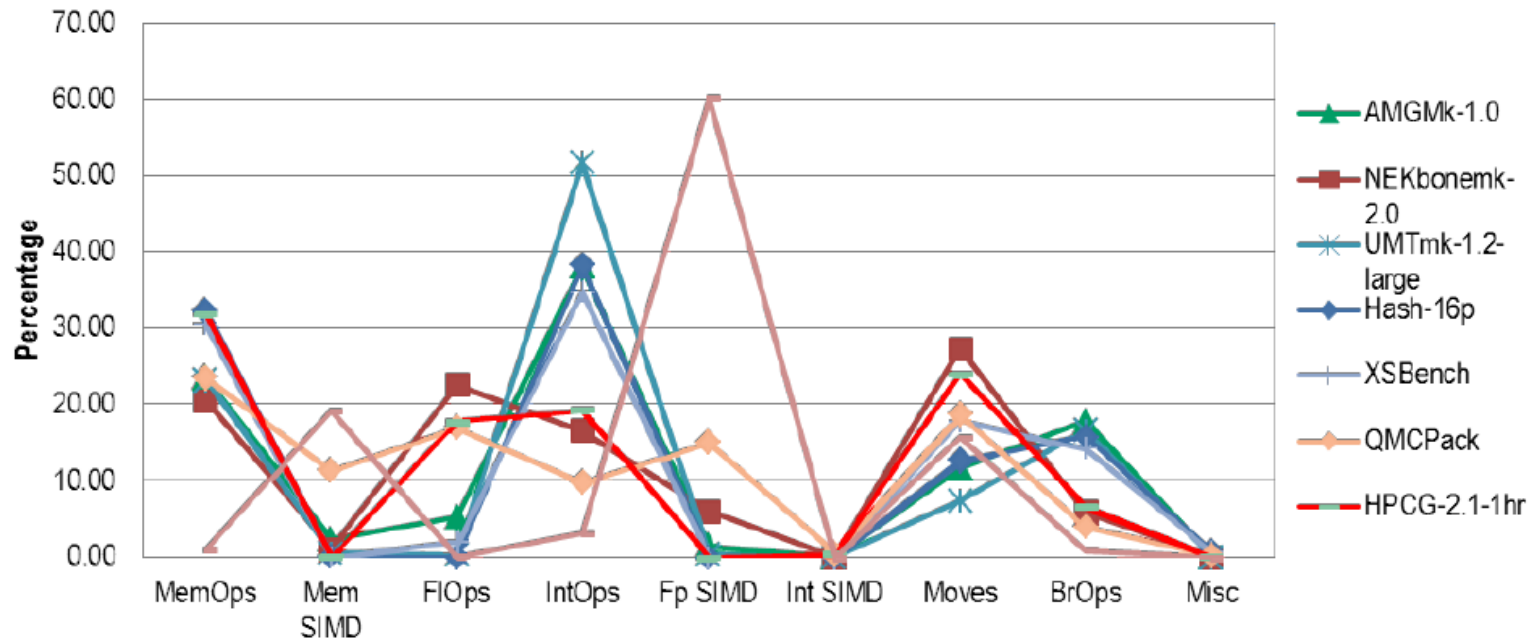
- One 'best configuration' on other architectures
- Major differences
 - Parallelism arrangement
 - Device-specific memory
 - Other arch optimizations

		Executed on		
		CUDA	GCN	MIC
Best Program version of	CUDA	100	84	65
	GCN	91	100	67
	MIC	58	68	100



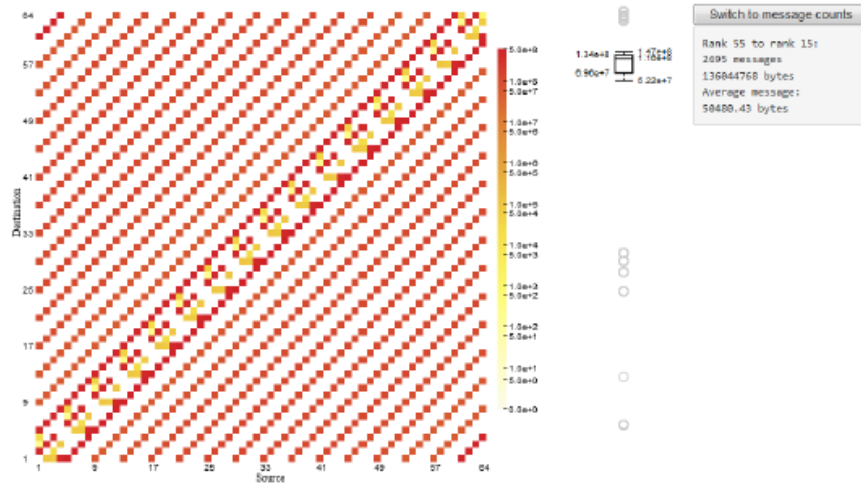
Application characteristics
(should) matter

Flops (and Integer SIMD) are Irrelevant

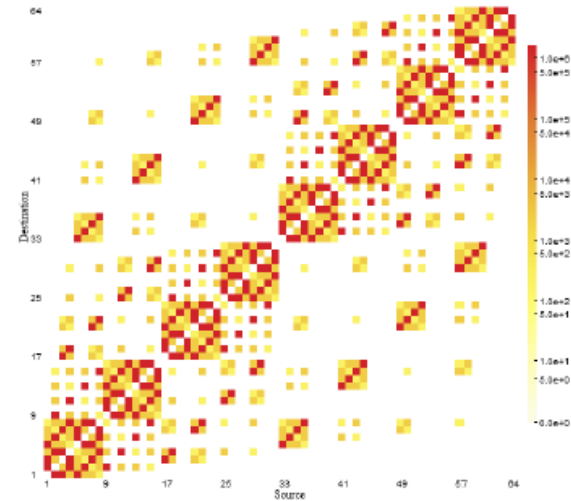


	MemOps	Mem SIMD	FIOps	IntOps	Fp SIMD	Int SIMD	Moves	BrOps	Misc
AMGMk-1.0	23.58	2.25	5.16	38.10	1.45	0.00	11.65	17.81	0.00
NEKbonemk-2.0	20.55	1.09	22.61	16.53	5.94	0.00	27.36	5.93	0.00
UMTmk-1.2-large	23.22	0.50	0.24	51.69	0.37	0.03	7.33	16.62	0.01
Hash-16p	32.38	0.08	0.00	38.31	0.00	0.02	12.63	15.92	0.64
XSBench	30.48	0.00	2.02	34.75	0.00	0.00	17.82	14.27	0.66
QMCPack	23.50	11.34	16.92	9.76	15.01	0.56	18.70	3.94	0.28
HPL	0.9	19.2	0.1	3.1	60.2	0	15.7	0.8	0
HPCG-2.1-1hr	31.85	0.040	17.76	19.23	0.002	0.40	24.034	6.62	0.036

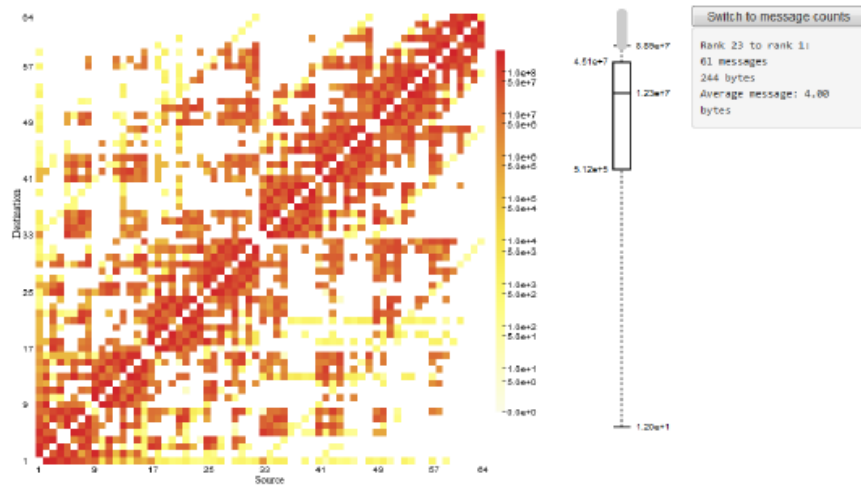
Communication patterns do exhibit structure at scale



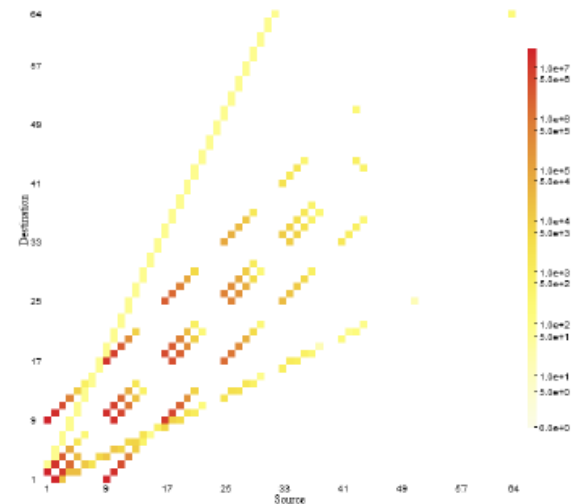
(c) HPL communication volume with boxplot and mouse-over data



(d) Multigrid_C communication volume



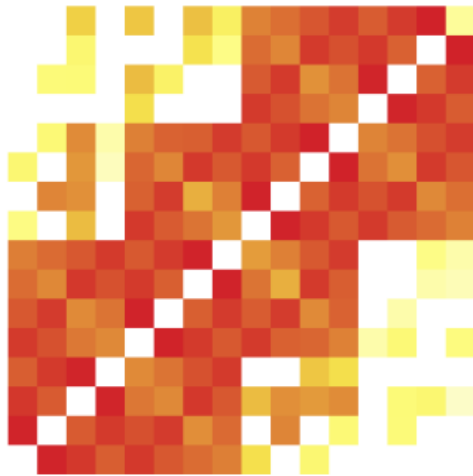
(e) miniAMR:expanding-sphere communication volume with boxplot and mouse-over data



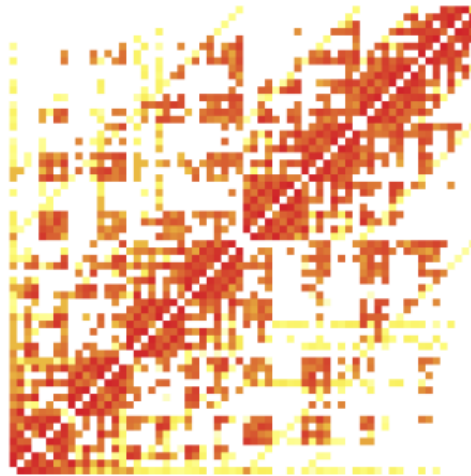
(f) MCB communication volume

Challenges of Input Dependent Applications

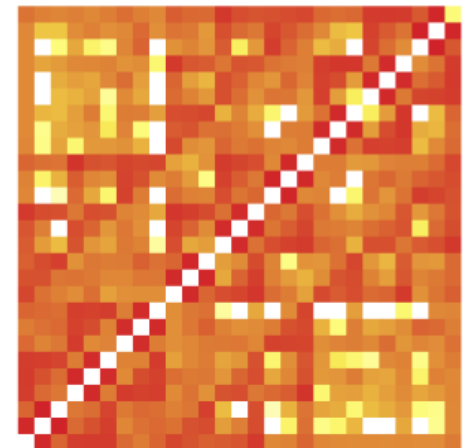
- E.g., Adaptive Mesh Refinement (AMR)
- Exactly the same code but different input problem
- We need new ways to capture/quantify this behavior



miniAMR - two-spheres-16p

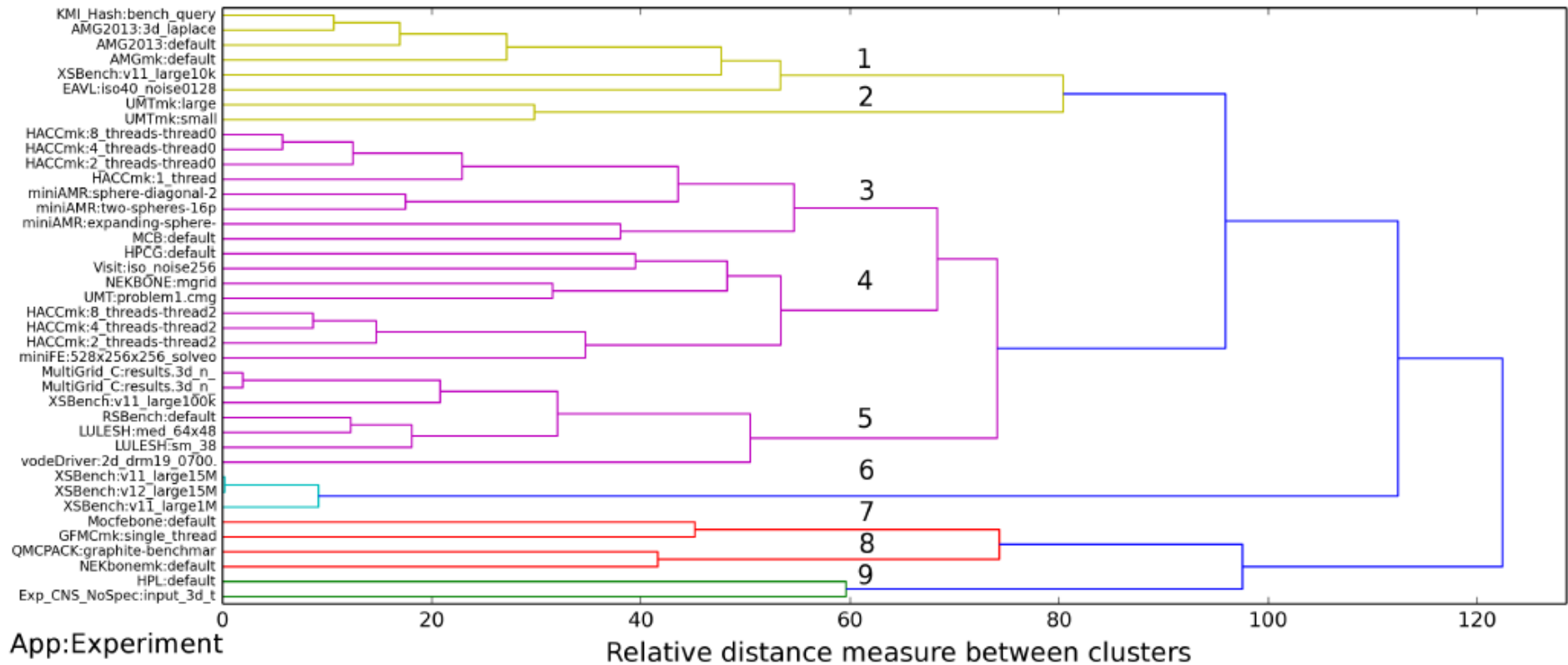


miniAMR - expanding-sphere-64p



miniAMR - sphere-diagonal-27p

Identify similarities across apps, benchmarks, proxies



Clustering on instruction mix measurements

Summary

- New and Improved Memory systems are the next Big Thing
- Heterogeneous computing is here to stay
- Application characteristics (should) matter

Contributors and Recent Sponsors

- Future Technologies Group: <http://ft.ornl.gov>
 - Publications: <https://ft.ornl.gov/publications>
- Department of Energy Office of Science
 - Vancouver Project: <https://ft.ornl.gov/trac/vancouver>
 - Blackcomb Project: <https://ft.ornl.gov/trac/blackcomb>
 - ExMatEx Codesign Center: <http://codesign.lanl.gov>
 - Cesar Codesign Center: <http://cesar.mcs.anl.gov/>
 - SciDAC: SUPER, SDAV <http://science.energy.gov/ascr/research/scidac/scidac-institutes/>
 - CS Efforts: <http://science.energy.gov/ascr/research/computer-science/>
- DOE 'Application' offices
- National Science Foundation Keeneland Project: <http://keeneland.gatech.edu>
- NVIDIA CUDA Center of Excellence at Georgia Tech
- Other sponsors
 - ORNL LDRD, NIH, AFRL, DoD
 - DARPA (HPCS, UHPC, AACE)

Q & A

More info: vetter@computer.org



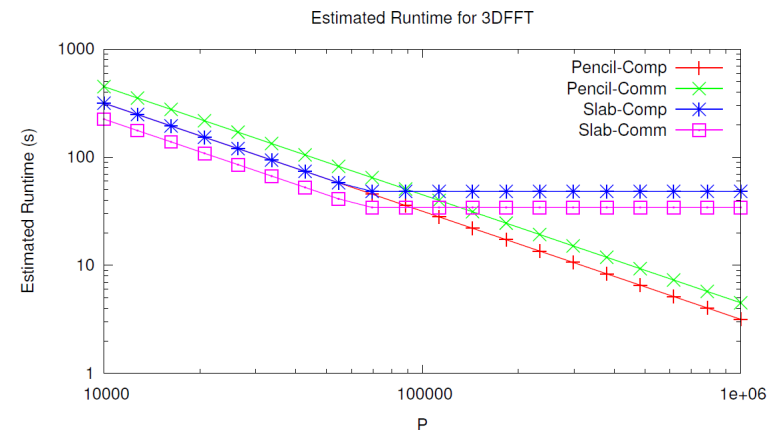
Ignore performance
prediction at your own risk

Aspen – Design Goals


- Abstract Scalable Performance Engineering Notation
 - Create a deployable, extensible, and highly semantic representation for analytical performance models
 - Design and implement a new language for analytical performance modeling
 - Use the language to create machine-independent models for important applications and kernels
- Models are composable

```
1 kernel localFFT {  
2   exposes parallelism [n^2]  
3   requires flops [5 * n * log2(n)] as dp,  
      complex, simd  
4   requires loads [a * n * max(1, log(n)/  
      log(Z)) * wordSize] from fftVolume  
5 }
```

Listing 2. Aspen statements for the local 1D FFTs



K. Spafford and J.S. Vetter, "Aspen: A Domain Specific Language for Performance Modeling," in *SC12: ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis*, 2012

 **jsmeredith** on Sep 20, 2013 adding models

1 contributor

336 lines (288 sloc) | 9.213 kb

Raw

Blame

History

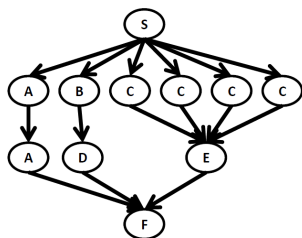


```

1  //
2  // lulesh.aspen
3  //
4  // An ASPEN application model for the LULESH 1.01 challenge problem. Based
5  // on the CUDA version of the source code found at:
6  // https://computation.llnl.gov/casc/ShockHydro/
7  //
8  param nTimeSteps = 1495
9
10 // Information about domain
11 param edgeElems = 45
12 param edgeNodes = edgeElems + 1
13
14 param numElems = edgeElems^3
15 param numNodes = edgeNodes^3
16
17 // Double precision
18 param wordSize = 8
19
20 // Element data
21 data mNodeList as Array(numElems, wordSize)
22 data mMatElemList as Array(numElems, wordSize)
23 data mNodeList as Array(8 * numElems, wordSize) // 8 nodes per element
24 data mIxm as Array(numElems, wordSize)
25 data mIxp as Array(numElems, wordSize)
26 data mletam as Array(numElems, wordSize)
27 data mletap as Array(numElems, wordSize)
28 data mzetam as Array(numElems, wordSize)
29 data mzetap as Array(numElems, wordSize)
30 data melemBC as Array(numElems, wordSize)
31 data mE as Array(numElems, wordSize)
32 data mP as Array(numElems, wordSize)

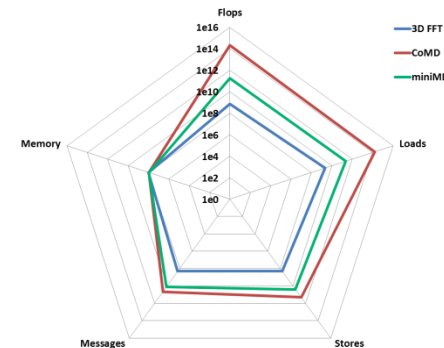
```

Aspen Design Flow



Representation in Aspen

- Modular
- Sharable
- Composable
- Reflects prog structure



Existing models for MD, UHPC CP 1, Lulesh, 3D FFT, CoMD, VPFFT, ...

Creation

- Static analysis
- Historical
- Empirical
- Manual



Use

- Drive simulators
- Feedback to runtime systems
- Design space optimization
- Interactive tools for graphs, queries

Informing Runtime Optimization

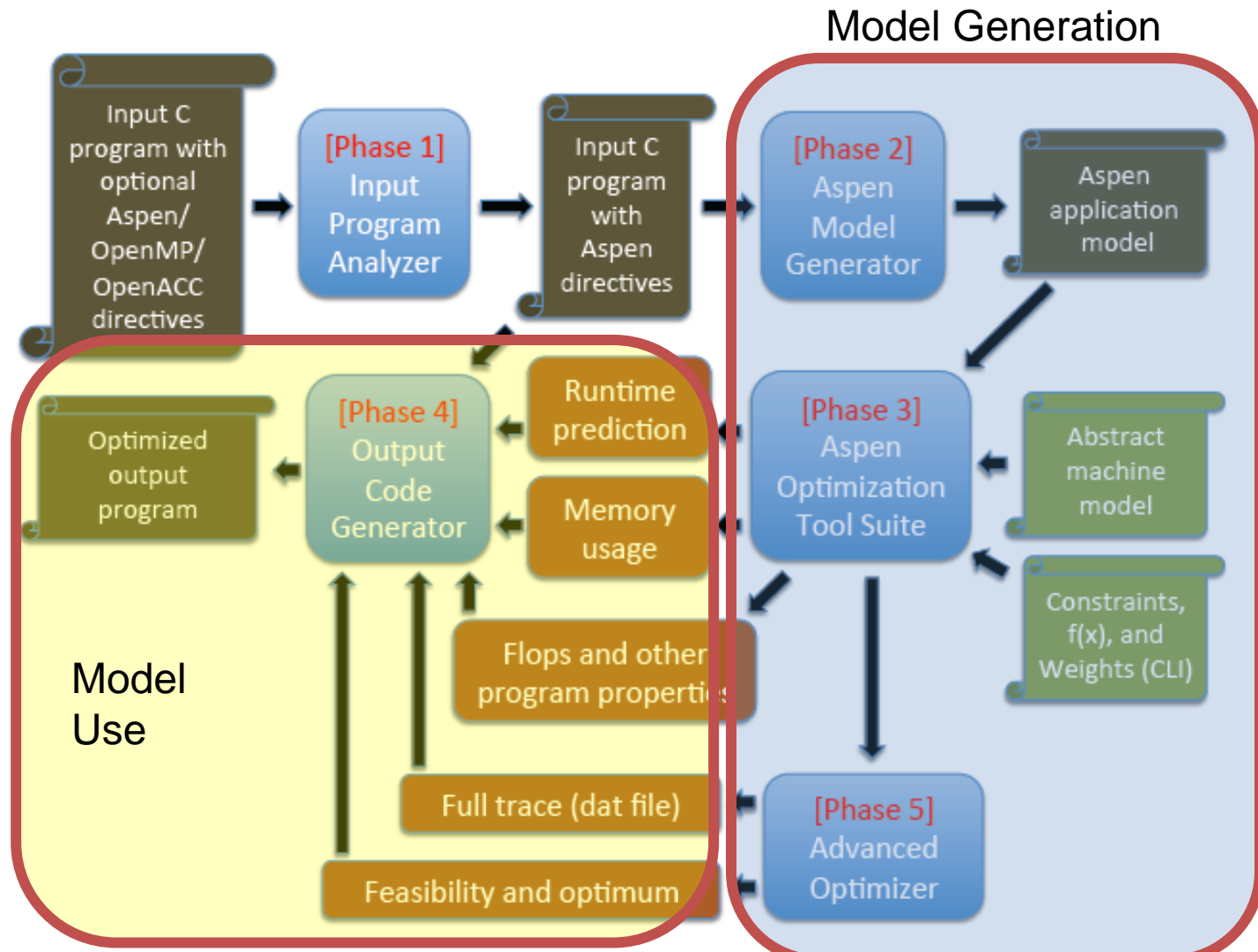


Fig. 1: Process diagram of the overall design exploration workflow.

Aspen Performance Models Drive Runtime Decisions

Listing 5: Input MM code modified to selectively offload OpenACC compute regions to GPUs based on the Aspen prediction

```
1  extern int HI_aspenpredict(double N); int N = 1024;
2  void matmul(float * A, float * B, float * C) { int i,j,k;
3  #pragma acc kernels loop if(HI_aspenpredict((double)N))
4  gang copyin(N, B[0:N*N], C[0:N*N]) copyout(A[0:N*N])
5    for (i=0; i<N; i ++ ) {
6    #pragma acc loop worker
7      for (j=0; j<N; j ++ ) { ... }
8    }; return ;
9  } //end of matmul()
10 int main() { ... }
```

LULESH – runtime optimizations

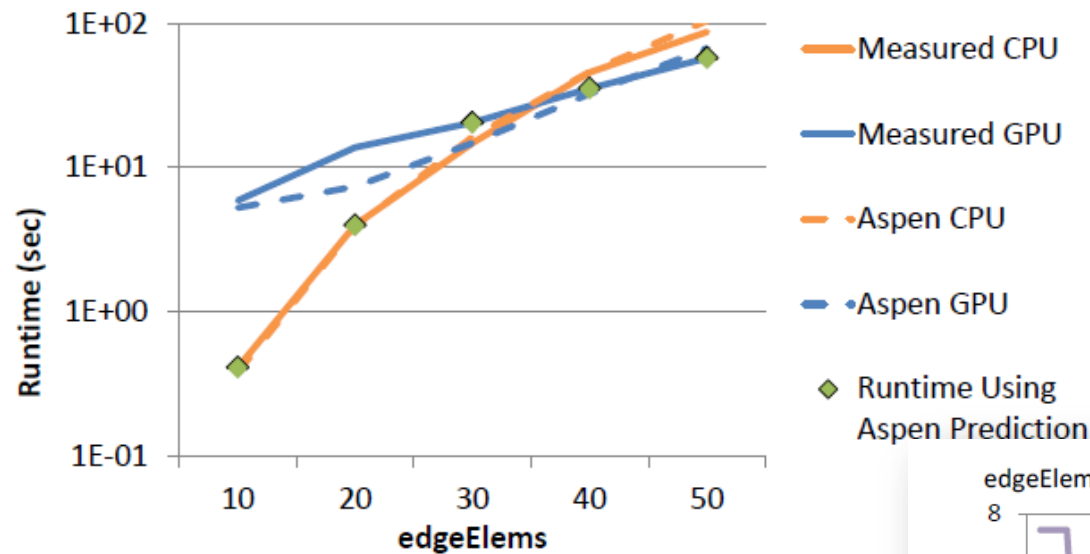


Fig. 7: Measured and predicted runtime of the entire program on CPU and GPU, including measured runtime of the automatically predicted optimal target device and

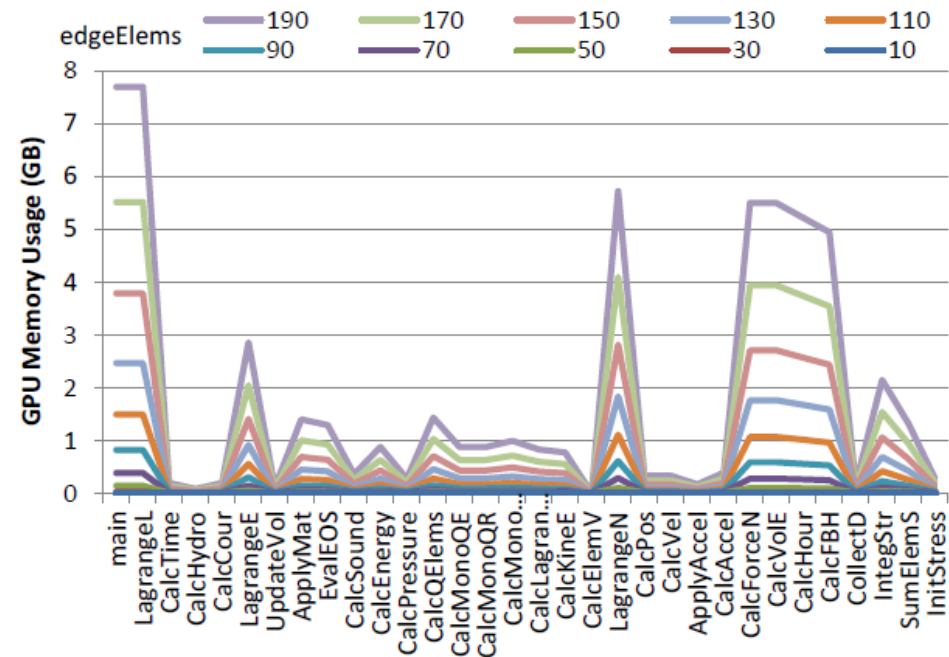


Fig. 8: GPU Memory Usage of each Function in LULESH, where the memory usage of a function is inclusive; value for a parent function includes data accessed by its child functions in the call graph.