

Exploring Emerging Technologies in the Extreme Scale HPC Co-Design Space with Holistic Performance Modeling

Jeffrey S. Vetter

Jeremy Meredith

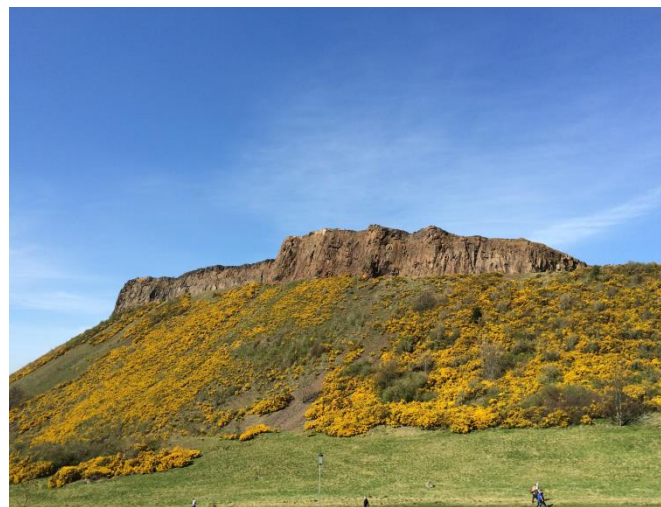
Presented to
Exascale Applications and Software
Conference (EASC)

EPCC/University of Edinburgh

22 Apr 2015

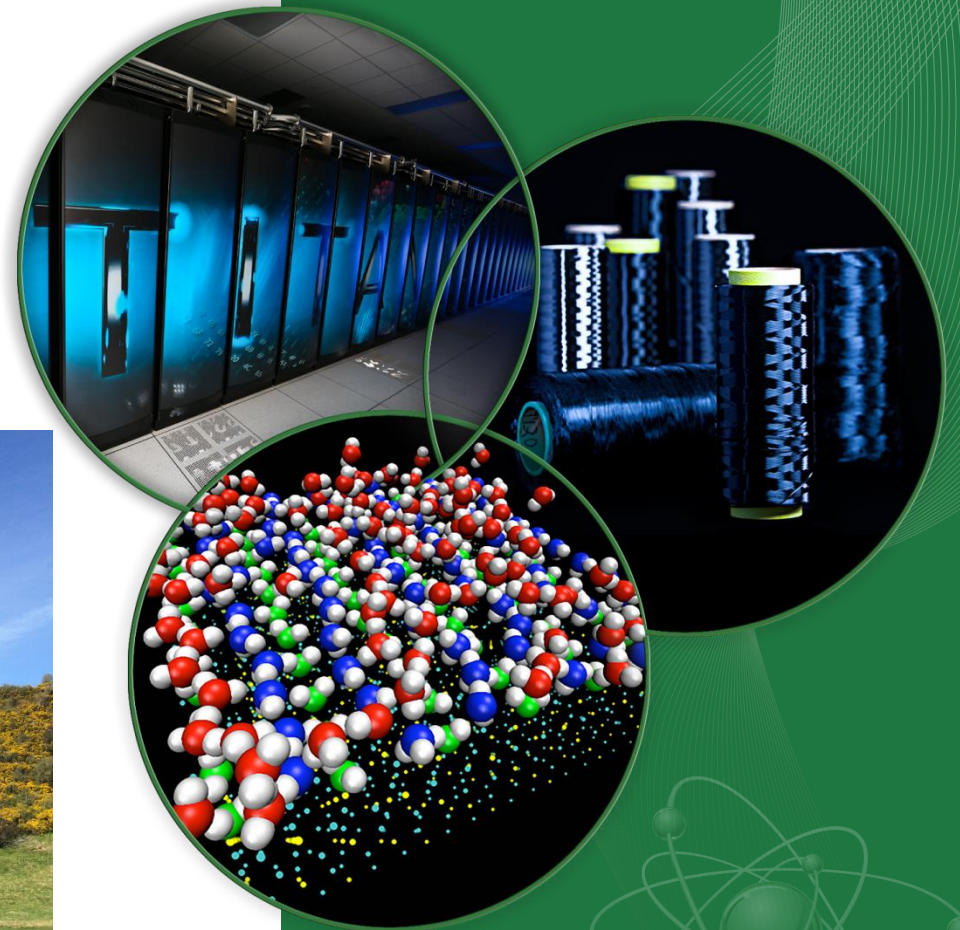
OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

ORNL is managed by UT-Battelle
for the US Department of Energy



Georgia Tech  **College of Computing**
Computational Science and Engineering

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



 **OAK RIDGE**
National Laboratory



*****WHISKY LIST*****

WHISKY REGIONS & TYPE OF WHISKY

L - Lowland	C - Campbell	H - Highlands	IS - Islands
IN - India	Town	I - Islay	J - Japanese
S - Speyside	B - Blend	NZ - New Zealand	W - Welsh



LOC	MALTS	Age	ABV	£	LOC	MALTS	Age	ABV	£
H Aberfeldy 12 YO		12	40.0%	£3.90	I Caol Ila 12 YO		12	43.0%	£5.50
H Aberfeldy 21 YO		21	40.0%	£9.90	I Caol Ila 18 YO		18	43.0%	£8.50
S Aberfeldy 18 YO		18	40.0%	£3.90	I Caol Ila Cask Strength		NL	50.0%	£8.80
S Aberfeldy 12 YO		12	40.0%	£4.50	I Caol Ila Distillers Edition		NL	43.0%	£7.50
S Aberfeldy 18 YO		18	43.0%	£6.50	I Caol Ila Moch		NL	43.0%	£5.80
S Aberfeldy 18 YO		18	43.0%	£9.90	I Caol Ila Unpeated Stitches R.		NL	59.6%	£22.00
S Aberfeldy a' bunadh (CS)		NL	60.2%	£5.90	I Caol Ila 25 YO		25	43.0%	£4.00
IN Amrut Sherry		NL	57.1%	£6.20	I Cardhu 12 YO		12	46.0%	£4.90
H Anenoc 12 YO		12	40.0%	£4.50	H Clynelish 14 YO		14	46.0%	£6.50
H Anenoc 18 YO		18	46.0%	£7.90	H Clynelish Dist Ed Oloroso		NL	46.0%	£5.50
I Ardbeg 10 YO		10	46.0%	£5.50	B Compass Box Spice Tree		NL	40.0%	£5.50
I Ardbeg Aulverdes		NL	49.9%	£9.90	S Cragganmore 12 YO		NL	40.0%	£7.50
I Ardbeg Corryvreckan		10	57.1%	£8.90	S Cragganmore Dist Ed		NL	40.0%	£5.90
I Ardbeg Uigedail		12	54.2%	£7.90	H Cu Dhubh Black		NL	40.0%	£5.50
H Ardmore Traditional Cask		NL	46.0%	£6.50	H Dalmore 12 YO		15	40.0%	£6.90
H Ardmore 1993		16	40.0%	£5.50	H Dalmore 18 YO		18	43.0%	£12.00
IS Arran 100% Proof		NL	57.0%	£4.50	H Dalmore 18 YO		NL	40.0%	£15.00
IS Arran 10 YO		10	43.0%	£4.20	H Dalmore Alexander III		NL	44.0%	£8.90
IS Arran 14 YO		14	46.0%	£5.90	H Dalmore Cigar Malt		25	42.0%	£45.00
IS Arran Cask Strength		12	54.1%	£6.50	H Dalmore 25 YO		25	43.0%	£5.50
L Auchentoshan Classic		NL	40.0%	£3.90	H Dalwhinnie 15 YO		15	43.0%	£6.90
L Auchentoshan 12 YO		12	40.0%	£4.50	H Dalwhinnie Distillers Edition		NL	43.0%	£6.90
L Auchentoshan 18 YO		18	43.0%	£8.90	H Deanston 12 YO		12	46.3%	£4.90
L Auchentoshan 3 Wood		NL	43.0%	£5.90	H Edradour 10 YO		10	40.0%	£5.50
S Auchrosk 10 YO		10	43.0%	£4.80	H Edradour Port Finish		13	55.7%	£6.90
H Balblair 1996		V	46.0%	£19.00	H Fettercairn Flor		NL	42.0%	£5.50
H Balblair 1975		18	43.0%	£11.00	S Glen Burgie 10 YO		12	43.0%	£3.90
H Balblair 1990		10	43.0%	£6.50	S Glen Elgin 12 YO		12	43.0%	£4.90
H Balblair 2003		12	43.0%	£3.90	H Glen Deveron 12 YO		12	40.0%	£4.90
S Balmenach (CC)		12	40.0%	£5.50	H Glen Garloch Founders Res.		NL	48.0%	£5.40
S Balvenie Doublewood 12 YO		12	40.0%	£5.50	H Glen Garloch 12 YO		10	40.0%	£4.50
S Balvenie Signature		12	40.0%	£5.50	S Glen Grant 10 YO		16	43.0%	£6.50
S Balvenie Single Barrel		15	47.8%	£6.50	S Glen Grant 16 YO		27	43.0%	£7.90
S Balvenie Rum Cask		14	43.0%	£5.90	S Glen Mhor 1980		16	40.0%	£4.90
S Balvenie Doublewood 17 YO		17	43.0%	£8.90	S Glen Moray 16 YO		10	46.0%	£4.50
S Balvenie 30 YO		30	47.3%	£35.00	C Glen Scotia 10 YO		10	46.0%	£7.50
S Banff 1976 (CC)		34	40.0%	£10.90	C Glen Scotia 18 YO		18	40.0%	£4.90
H Ben Nevis 10 YO		10	46.0%	£3.90	C Glen Scotia 12 YO		12	40.0%	£4.90
S Benriach 12 YO		12	43.0%	£4.00	C Glen Scotia Cask Strength		15	59.9%	£5.80
S Benriach Pated 16 YO		10	40.0%	£4.00	S Glen Spey 12 YO		12	43.0%	£4.90
S Benriach Dark Rum Finish		15	46.0%	£6.90	H Glencadam 10 YO		10	46.0%	£4.00
S Benriach Tammy Port Finish		16	46.0%	£5.50	H Glencadam Oloroso		14	46.0%	£5.50
S Benriach Solistica		17	50.0%	£6.50	H Glencadam 15 YO		15	40.0%	£5.40
S Benriach Bernie Moss		NL	48.0%	£3.80	S Glendullan 12 YO		12	43.0%	£6.00
S Benromach 10 YO		10	43.0%	£4.20	H Glendronach 12 YO		12	40.0%	£4.90
S Benromach Organic		NL	43.0%	£5.50	H Glendronach 15 YO		15	46.0%	£5.90
S Benromach Peat Smoke		NL	46.0%	£4.60	H Glendronach 18 YO		18	46.0%	£8.50
S Benromach Origins		NL	50.0%	£5.40	H Glendronach 21 YO		21	48.0%	£12.00
H Blair Athol 12 YO		12	43.0%	£5.90	H Glendronach Cask Strength		NL	54.8%	£7.50
H Blair Athol 1997 (CC)		15	43.0%	£4.90	S Glenfarclas 10 YO		10	40.0%	£4.90
I Bowmore Tempest		10	53.5%	£5.90	S Glenfarclas 15 YO		15	40.0%	£5.90
I Bowmore Original		12	40.0%	£5.50	S Glenfarclas 106 (CS)		10	60.0%	£5.90
I Bowmore 18 YO		18	43.0%	£9.50	S Glenfarclas 21 YO		21	43.0%	£9.00
I Bowmore Darkest		15	40.0%	£7.90	S Glenfiddich 12 YO		12	40.0%	£4.90
I Bowmore 25 YO		25	43.0%	£35.00	S Glenfiddich 18 YO		18	40.0%	£12.00
I Bruichladdich Scottish Barley		NL	50.0%	£5.50	S Glenfiddich Rum 21 YO		21	40.0%	£15.00
I Bruichladdich Port Charlotte		NL	50.0%	£8.50	S Glenfiddich 30 YO		30	40.0%	£29.00
I Bruichladdich Octomore 6.1		5	57.0%	£11.00	S Glenfiddich Solera 15 YO		15	40.0%	£6.00
I Bruichladdich Peat		NL	46.0%	£4.90	S Glenfiddich Rich Oak		14	40.0%	£6.50
I Bruichladdich Redder Still		22	50.4%	£20.00	H Glenglassaugh Revival		NL	46.0%	£4.60
I Bruichladdich Islay Barley		NL	50.0%	£5.90	H Glengoyne 10 YO		10	40.0%	£4.90
I Bunnahabhain 12 YO		12	46.3%	£4.90	H Glengoyne Cask Strength		NL	58.7%	£6.50
I Bunnahabhain 18 YO		18	46.3%	£8.50	H Glengoyne 15 YO		15	43.0%	£6.50
I Bunnahabhain 25 YO		25	43.0%	£27.00	H Glengoyne 18 YO		18	43.0%	£9.50
I Bunnahabhain Toitich		NL	46.0%	£6.90	S Glen Keith (CC)		12	46.0%	£5.20



*****WHISKY LIST*****

WHISKY REGIONS & TYPE OF WHISKY

L - Lowland	C - Campbell	H - Highlands	IS - Islands
IN - India	Town	I - Islay	J - Japanese
S - Speyside	B - Blend	NZ - New Zealand	W - Welsh

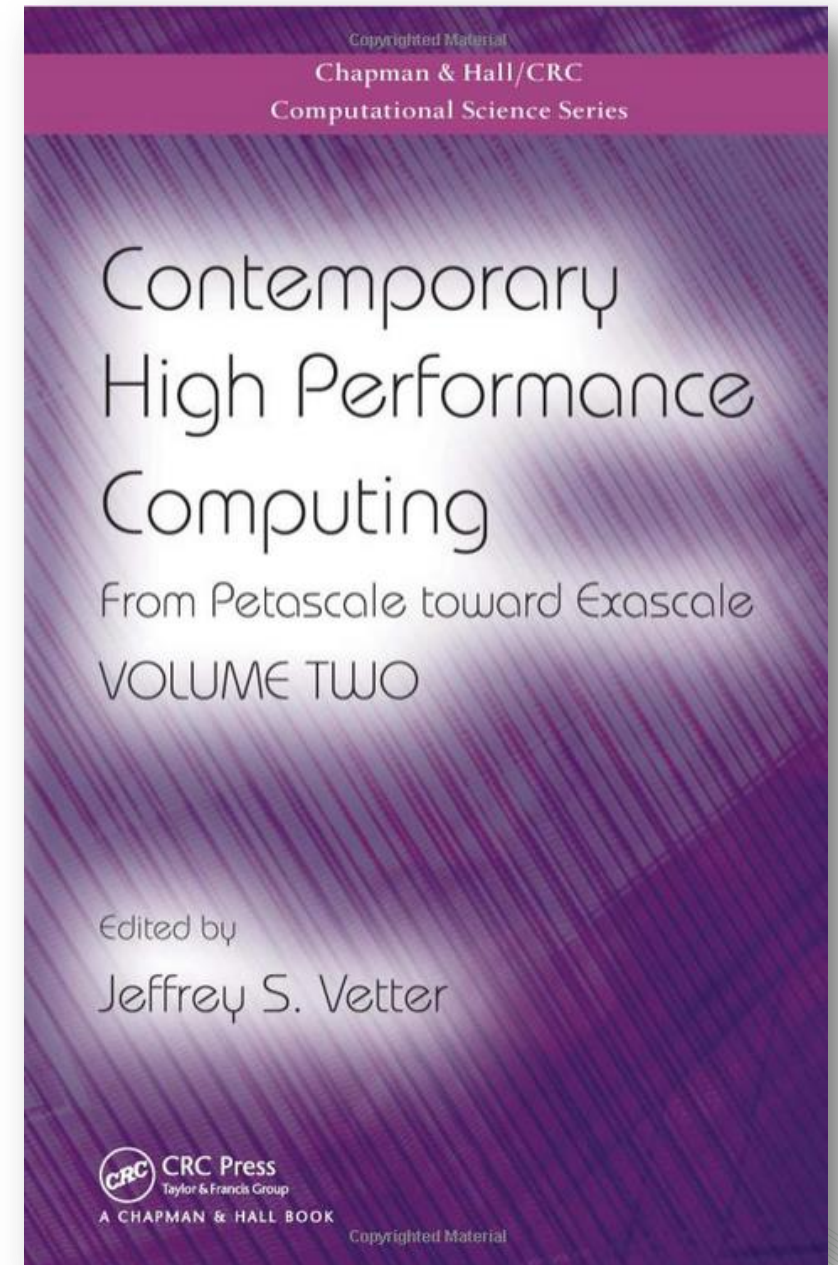


LOC	MALTS	Age	ABV	£	LOC	MALTS	Age	ABV	£
L Glenkinchie 12 YO		12	43.0%	£5.50	S Macallan Sherry 18 YO		18	43.0%	£9.50
L Glenkinchie Distillers Edition		NL	43.0%	£6.50	S Macallan 1972		V	43.0%	£22.00
L Glenkinchie 20 YO (CS)		20	56.4%	£14.00	S Macduff 1997 (CC)		14	40.0%	£4.70
S Glenlivet 12 YO		12	40.0%	£4.90	B Mackinnays Shackleton		NL	47.3%	£14.00
S Glenlivet 18 YO		18	43.0%	£6.50	S Mannochmore 12 YO		12	43.0%	£5.90
S Glenlivet Nadurra Cask (CS)		16	57.7%	£6.90	S Milton Duff 18 YO		15	43.0%	£5.80
H Glenmorangie Original		10	40.0%	£4.90	S Mortlach 15 YO		15	43.0%	£5.90
H Glenmorangie Nectar D'Or		15	46.0%	£6.50	S Mortlach 21 YO		21	43.0%	£9.50
H Glenmorangie Lasanta		12	46.0%	£5.90	S Mortlach Rare and Old		NL	43.4%	£7.50
H Glenmorangie Quinta Ruban		NL	46.0%	£5.50	NZ N.Z. Dunedin Doublewood		10	40.0%	£7.00
H Glenmorangie 18 YO		18	43.0%	£9.90	H Oban 14 YO		14	43.0%	£5.90
H Glenmorangie Signet		NL	46.0%	£14.90	H Oban Distillers Edition		NL	43.0%	£7.50
H Glenmorangie 25 YO		25	43.0%	£30.00	H Old Pulteney 12 YO		12	40.0%	£4.40
S Glenrothes Reserve		NL	43.0%	£4.90	H Old Pulteney Cask Strength		15	60.5%	£6.50
S Glentauchers 1991		15	40.0%	£3.90	H Old Pulteney 17 YO		17	46.0%	£6.80
H Glenturret		10	40.0%	£3.90	H Old Pulteney 21 YO		21	46.0%	£12.00
C Hazelburn		8	46.0%	£6.50	W Penderyn		NL	46.0%	£4.00
IS Highland Park 12 YO		12	40.0%	£4.90	IS Robert Burns 12 YO		12	43.0%	£3.90
IS Highland Park 18 YO		18	43.0%	£12.00	H Royal Brackla 1996		16	46.0%	£4.50
IS Highland Park 21 YO		21	47.5%	£16.00	H Royal Lochnagar 12 YO		12	40.0%	£4.90
IS Highland Park 26 YO		25	46.1%	£25.00	H Royal Lochnagar Dist Ed		NL	40.0%	£5.80
IS Highland Park 30 YO		30	48.1%	£45.00	IS Scape 18 YO		16	40.0%	£6.50
IS Highland Park Cask Strength		11	58.2%	£6.50	IS Scape 25 YO		25	54.0%	£25.50
IS Highland Park Dark Origins		NL	46.8%	£7.50	S Singleton of Dufttown		12	43.0%	£4.90
S Imperial		16	43.0%	£5.50	I Smokehead		NL	43.0%	£4.80
S Inchgower		14	43.0%	£5.00	I Smokehead 18 YO		18	46.0%	£9.90
B Johnnie Walker Blue Label		NL	40.0%	£18.00	S Speyburn 10 YO		10	40.0%	£3.90
IS Jura 10 YO		10	40.0%	£4.90	C Springbank 10 YO		10	46.0%	£4.90
IS Jura 16 YO		16	40.0%	£5.90	C Springbank 15 YO		15	46.0%	£5.90
IS Jura Elixir		12	40.0%	£4.90	C Springbank 18 YO		18	46.0%	£6.90
IS Jura Superstition		NL	45.0%	£5.90	C Springbank Calvados		12	52.7%	£6.90
IS Jura Prophecy		NL	46.0%	£7.50	S Strathisla 12 YO		12	43.0%	£35.00
IS Jura 200th Anniversary		21	44.0%	£12.90	S Strathisla 1967		50	43.0%	£4.70
IS Jura Tasting		NL	44.0%	£8.90	S Strathmill 12 YO		12	46.8%	£5.50
I Kilchoman Machir Bay		NL	46.0%	£5.50	IS Talisker 10 YO		18	45.8%	£8.90
I Kilchoman 100% Islay		NL	50.0%	£7.50	IS Talisker 18 YO		12	45.8%	£6.50
I Kilchoman 2007		6	46.0%	£5.50	IS Talisker Distillers Edition		NL	57.0%	£7.50
C Kilkerran		NL	46.0%	£4.50	IS Talisker 57 North		NL	45.8%	£5.90
S Knockando		12	40.0%	£4.50	IS Talisker Storm		NL	45.8%	£7.50
IS Lagavulin Cask Strength		12	56.4%	£8.90	IS Talisker Port Rulghie		NL	45.8%	£7.50
I Lagavulin 16 YO		16	43.0%	£6.50	IS Talisker 30 YO		30	53.1%	£25.00
I Lagavulin Distillers Edition		NL	43.0%	£8.50	S Tamdhu 8 YO		8	43.0%	£4.00
I Laphroaig 10 YO		10	40.0%	£4.90	S Tamdhu 10 YO		10	40.0%	£4.50
I Laphroaig 18 YO		18	48.0%	£11.00	S Tamdhu 16 YO		NL	43.0%	£25.00
I Laphroaig Quarter Cask		15	48.0%	£5.90	S Teaninich 10 YO		10	43.0%	£4.90
I Laphroaig Triple Wood		NL	48.0%	£6.90	H Teaninich (CC)		15	46.0%	£4.90
IS Ledaig 10 YO		10	46.3%	£4.00	IS Tobermory		10	46.3%	£4.50
S Linkwood 15 YO		15	43.0%	£4.60	IS Tobermory 16 YO		15	46.3%	£6.50
S Linkwood 12 YO		12	43.0%	£5.00	S Tomatin 12 YO		12	40.0%	£3.80
S Longmorn 16 YO		16	48.0%	£5.90	S Tomatin 18 YO		18	46.0%	£5.50
C Longrow		NL	46.0%	£5.60	S Tomintoul 'Peaty Tang'		NL	40.0%	£4.50
C Longrow 18 YO		18	46.0%	£9.90	S Tormore (CC)		NL	43.0%	£4.20
C Longrow Red Port Cask		11	51.8%	£6.50	S Tormore 12 YO		12	40.0%	£3.80
S Macallan Gold		NL	40.0%	£4.90	H Tullibardine 1993		10	46.0%	£4.80
S Macallan Amber		NL	40.0%	£7.50	IR Tyrconnell		NL	40.0%	£3.80
S Macallan Sienna		NL	43.0%	£9.90	J Yamazaki 12 YO		12	43.0%	£8.50
S Macallan Ruby		NL	43.0%	£15.00	J Yoichi 10 YO		10	45.0%	£9.80
S Macallan 2003		9	40.0%	£5.50					

ARCHER in new book

- Vol 2:
 - **EPCC - ARCHER**
 - NERSC
 - NREL
 - NCAR
 - ZIB
 - RIKEN
 - KTH Royal Institute of Technology

<http://j.mp/chpc2015>



Overview

- Our community has major challenges in HPC as we move to extreme scale
 - Power, Performance, Resilience, Productivity
 - Major shifts in architectures, software, applications
 - Not just HPC: Most uncertainty in two decades
- New technologies emerging to address some of these challenges
 - Heterogeneous computing
 - Nonvolatile memory
- Consequently, we now have critical situations in
 - Portable programming models
 - Performance prediction for procurement, optimization, etc
- Aspen is a tool we have developed for performance prediction



Surveying the HPC Landscape: Today and Tomorrow

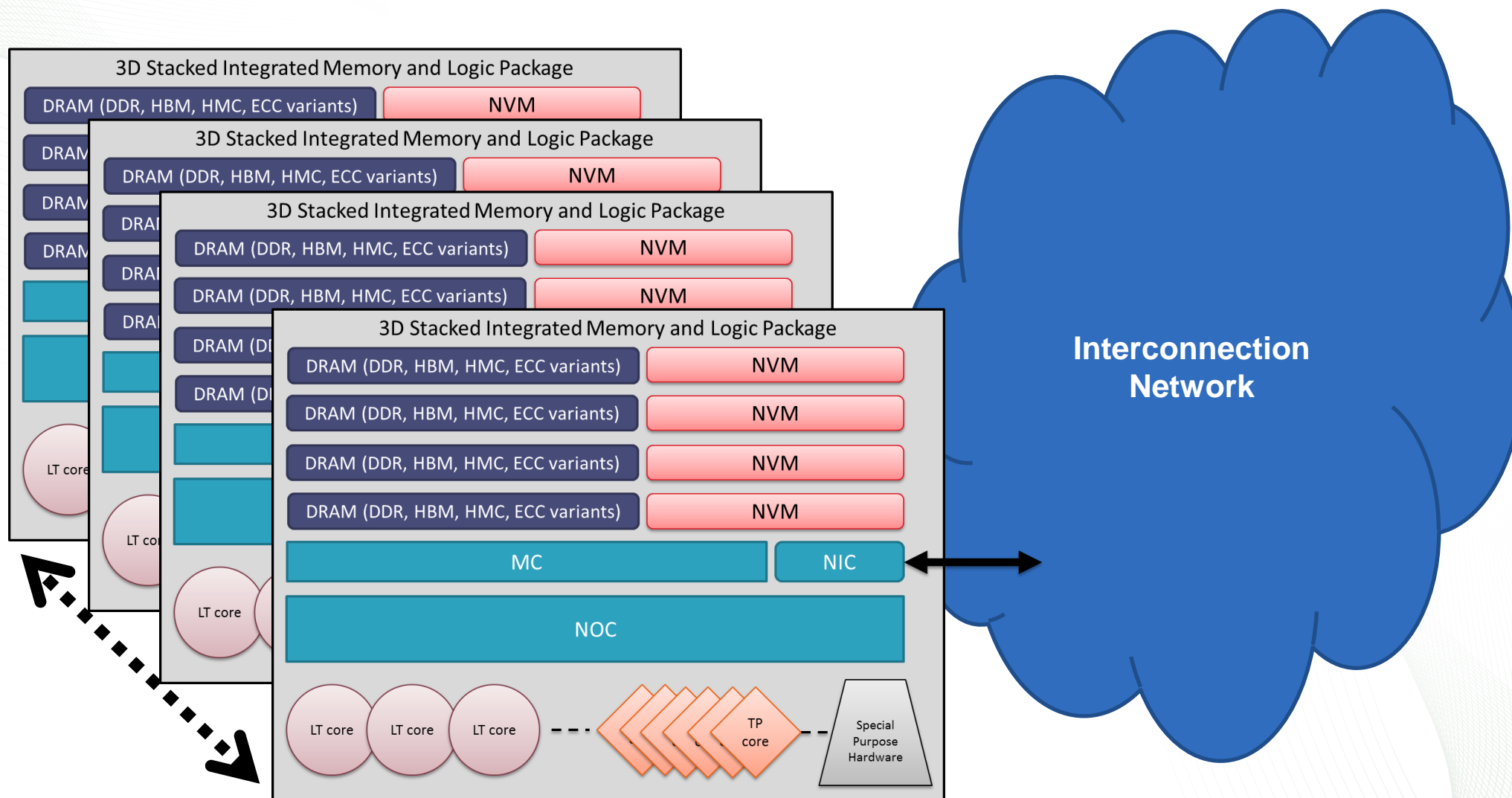
Notional Exascale Architecture Targets

(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 ??

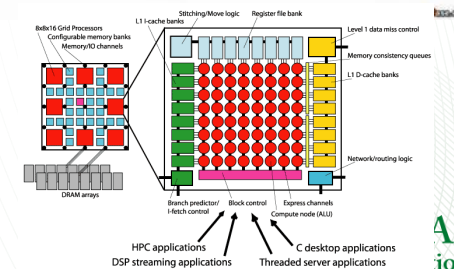
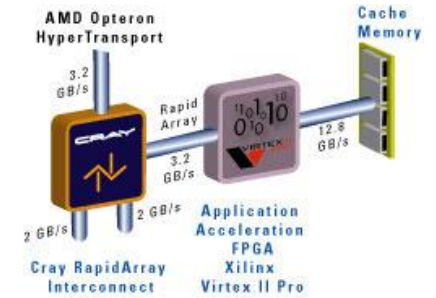
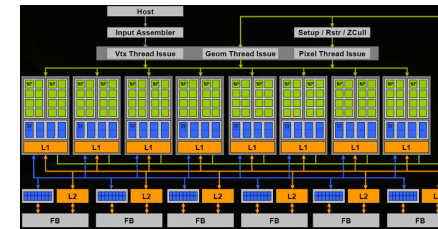
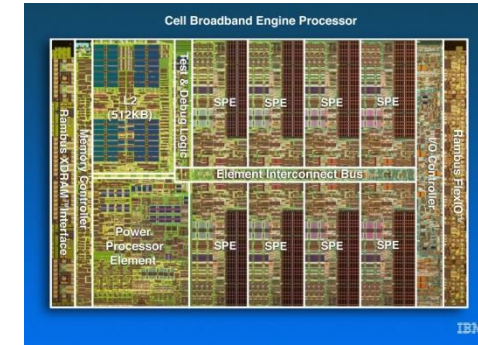
Notional Future Architecture



Earlier Experimental Computing Systems

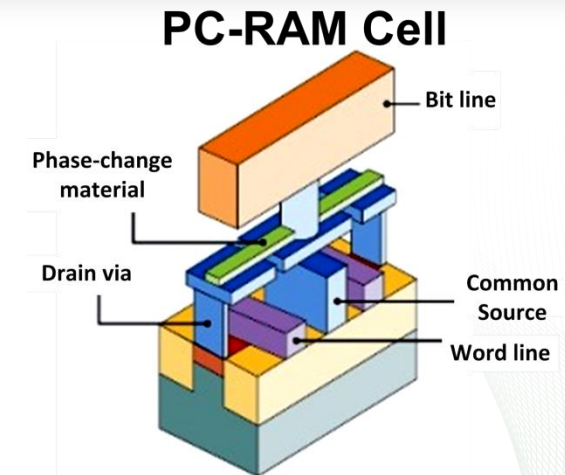
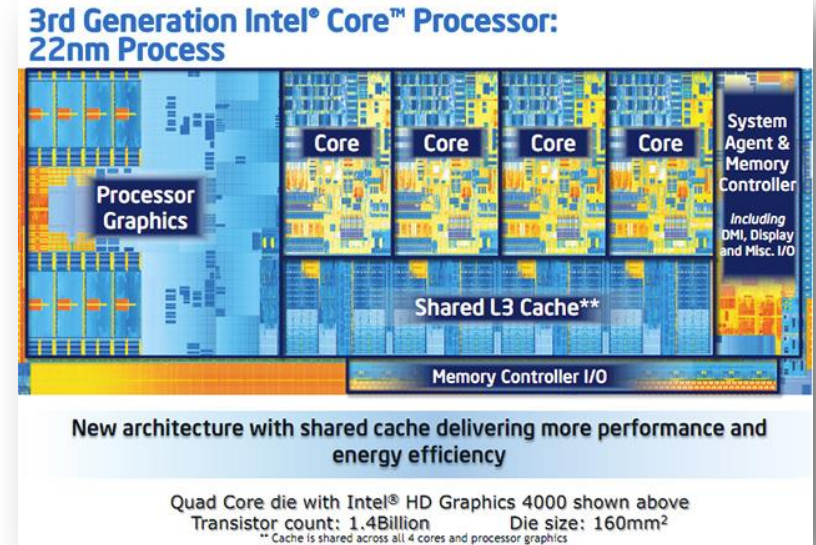
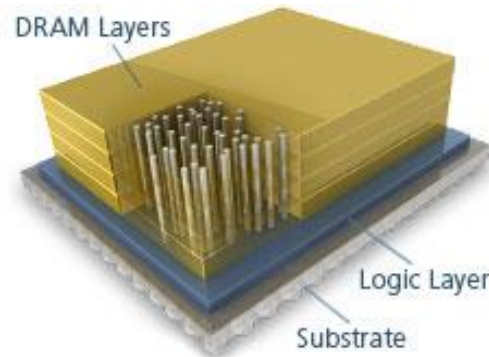
- The past decade has started the trend away from traditional 'simple' architectures
- Mainly driven by facilities costs and successful (sometimes heroic) application examples
- Examples
 - Cell, GPUs, FPGAs, SoCs, etc
- Many open questions
 - Understand technology challenges
 - Evaluate and prepare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



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

Recent announcements

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 - M

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
Comment

"SEATTLE" 64-BIT ARM SERVER PROCESSOR FIRST 28NM ARM

Nvidia Jetson TK1 mini supercomputer is up for pre-order

Will ship on 15 May

By Lee Bell

Fri May 02 2014, 11:11



computers attempt
self-driving cars.

Speaking at the GPU
Hsun Huang describ
can run, but at a sl

With a total perform
Raspberry Pi board
in the US - a numb
launched at CES in

"The Jetson TK1 als
comes with a whole
it could be classifi

Parameters are loa
recognises objects,

MarketWatch

PRESS RELEASE

Altera and IBM Unveil FPGA-accelerated POWER Systems with Coherent Shared Memory

By
Published: Nov 17, 2014 8:00 a.m. ET

f 8 t 13 8+ e

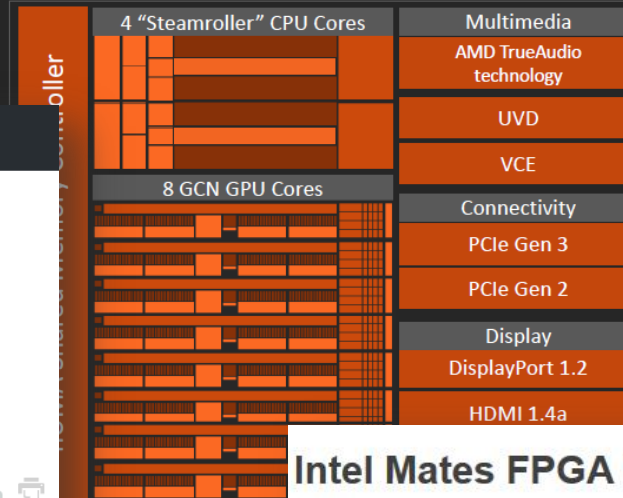
POWER8 Systems that Leverage Reprogrammable FPGA Accelerators Gain
Significant Improvements in System Performance, Efficiency and Flexibility

NEW ORLEANS, Nov. 17, 2014 /PRNewswire/ -- **SuperComputing 2014** -- Altera Corporation ALTR, +0.00% and IBM IBM, +0.00% today unveiled the industry's first FPGA-based acceleration platform that coherently connects an FPGA to a POWER8 CPU leveraging IBM's Coherent Accelerator Processor Interface (CAPI). The reconfigurable hardware accelerator features shared virtual memory between the FPGA and processor which significantly improves system performance, efficiency and flexibility in high-performance computing (HPC) and data center applications. Altera and IBM are presenting several POWER8 systems that are coherently accelerated using FPGAs at SuperComputing 2014.

Working together through the OpenPOWER Foundation, Altera and IBM are

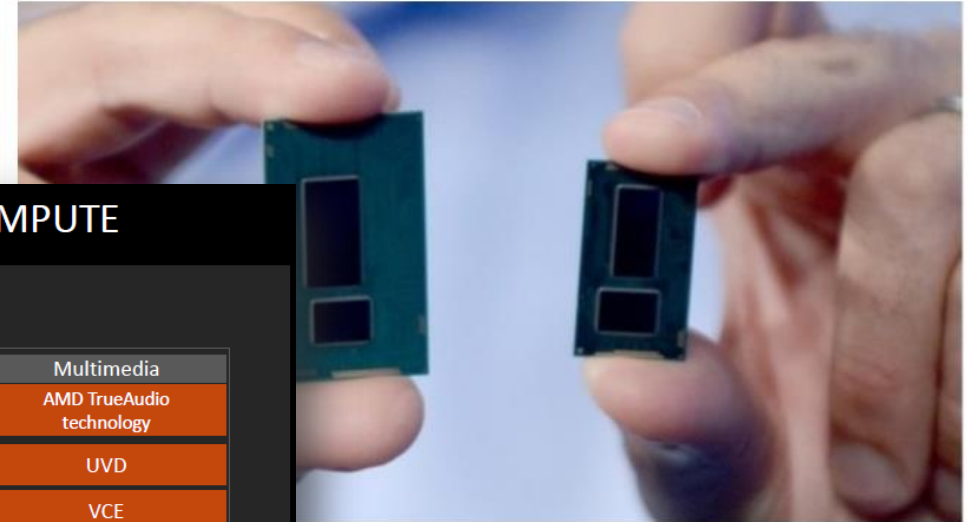
A-SERIES REDEFINES COMPUTE

Kaveri



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 microarchitecture, and while the CPU side of things isn't expected to change much, Broadwell's GPU looks

Intel Mates FPGA With Future Xeon Server Chip

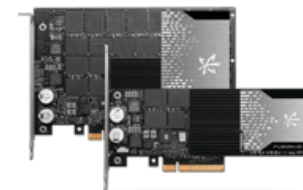
June 18, 2014 by Timothy Prickett Morgan



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.

RIDGE
National Laboratory

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

NO RATINGS
1 saves
LOGIN TO RATE

f Like 17 Tweet

LONDON — Samsung production of a 128 G multiple layers, and cl

The memory is based conventional floating g In the vertical arrange reliability between a fa conventional floating- in a [press release](#).

The technology is cap did not disclose how n vertical NAND, nor wh whether it had relaxed in 2D memory, which s

The company did say that the memory would provide improvements in performance and area ratio, and a V-NAND chip is suitable for a wide range of consumer and commercial applications including embedded NAND storage and solid-state drives.

The V-NAND component has the same memory capacity as a 128

designlines MEMORY

News & Analysis

3D NAND Transition: 15nm Process Technology Takes Shape

Gary Hilson

5/13/2014 08:15 AM EDT
5 comments

NO RATINGS
LOGIN TO RATE

f Like 15 Tweet 6 in Share 6 g+1 1

TORONTO — With 3D NAND unlikely to make economic sense until y partner Toshiba both ologies to produce NAND



Nelson said there is room to advance floating gates before moving

http://www.eetasia.com/STATIC/ARTICLE_IMAGES/201212/EEOL_2012DEC28_STOR_MFG_NT_01.jpg

Original URL: <http://www.theregister.com>

HP 100TB Memristor drives

Universal memory slow in com

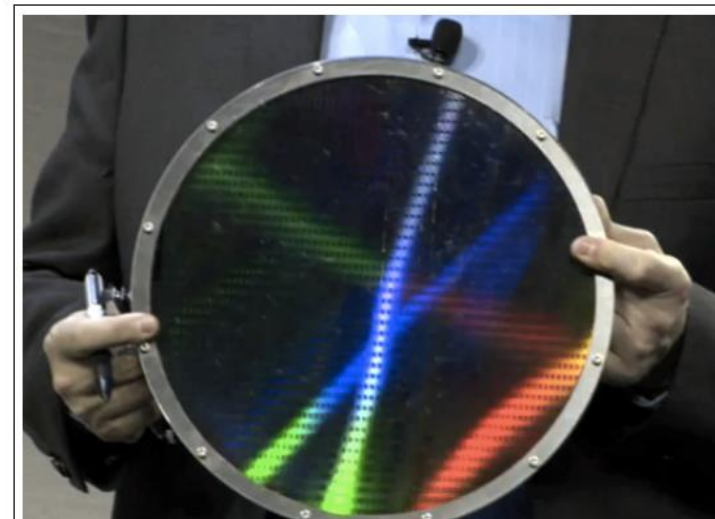
By Chris Mellor

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 have DRAM-like speed and better-than-NAND storage density. Fink claimed at an HP Discover event in Las Vegas that Memristor devices will be ready by the time flash NAND hits its limit in five years. He also showed off a Memristor wafer, adding that it could have a 1.5PB capacity by the end of the decade.



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									

Thinking back to 2009 projections, where is DOE in 2015?

System attributes	Today		CORAL	
Name	TITAN	MIRA	Summit	Aurora
System peak (PF)	27	10	150	180
Peak Power (MW)	9	4.8	10	13
Total system memory	710TB	768TB	2 PB DDR4 + HBM + 2.7 PB persistent memory	>7 PB High Bandwidth On-Package Memory, local Memory and Persistent Memory
Node performance (TF)	1.452	0.204	> 40	> 17 times Mira
Node processors	AMD Opteron Nvidia Kepler	64-bit PowerPC A2	Multiple IBM Power9 CPUs & multiple Nvidia Voltas GPUS	Intel Xeon Phi processors (codenamed Knights Hill)
System size (nodes)	18,688 nodes	49,152	>3,400 nodes	>50,000 nodes
System Interconnect	Gemini	5D Torus	Dual Rail EDR-IB	2nd generation Intel Omni-Path Architecture
File System	32 PB 1 TB/s, Lustre®	26 PB 300 GB/s GPFS™	120 PB 1 TB/s GPFS™	150 PB >1 TB/s Lustre®

Some ratios will be challenging to mitigate

System attributes	2001	2010	2014	"2015"		est 2018	Ratio of Summit to Titan	"2018"	
Name	Seaborg3	Jaguar	Titan			SUMMIT			
System peak	10 Tera	2 Peta	27	200		136	5.04	1 Exaflop/sec	
Power (MW)	0.8	6	9	15		10	1.11	20	
Node main memory (GB)			38			512	13.47		
System memory (PB)	0.006	0.3	0.7106	5		1.7408	2.45	32-64	
Node Persistent Memory (GB)						800			
System Persistent Memory (PB)						2.72	∞		
Node performance (TF)	0.024	0.125	1.4	0.5	7	40	28.57	1	10
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)	*POWER9s + *VOLTAs		O(1,000)	O(10,000)
System size (nodes)	416	18700	18700	50000	5000	3400	0.18	1000000	100000
Total Node Interconnect BW		1.5 GB/s		150 GB/sec	1 TB/sec			250 GB/sec	2 TB/sec
injection bandwidth per node (GB/s)			6.4			23	3.59		
File system capacity (PB)			32			120	3.75		
File system bandwidth (TB/s)			1			1	1.00		
MTTI		day		O(1 day)				O(1 day)	

Observations about these trends

- Aside from all the interesting technical questions for computer scientists...

DANGER



HALT!



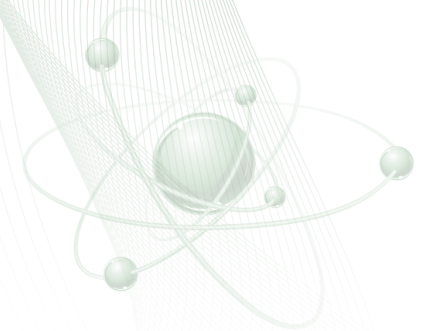
DEADLY FORCE
IS AUTHORIZED
BEYOND THIS POINT

10 Code of Federal Regulations 1047 and Department of Energy Manual 470.4-3

Observations about these trends (2)

1. For the success of HPC, we need to be very careful at this point
2. Complexity is everyone's enemy!
3. Performance portable programming models should be mandatory on all current and future architectures
 1. Increasingly, apps teams are spending time porting to new architectures rather than doing science
4. Performance prediction techniques and tools are critical
 1. Previously, a poor (procurement, optimization, facility) decision could cost 30%; now it could be 10x!
5. And then there is power consumption, reliability, etc

Holistic Performance Modeling for Extreme-Scale HPC



Prediction Techniques Ranked

	Speed	Ease	Flexibility	Accuracy	Scalability
Ad-hoc Analytical Models	1	3	2	4	1
Structured Analytical Models	1	2	1	4	1
Simulation – Functional	3	2	2	3	3
Simulation – Cycle Accurate	4	2	2	2	4
Hardware Emulation (FPGA)	3	3	3	2	3
Similar hardware measurement	2	1	4	2	2
Node Prototype	2	1	4	1	4
Prototype at Scale	2	1	4	1	2
Final System	-	-	-	-	-

Prediction Techniques Ranked

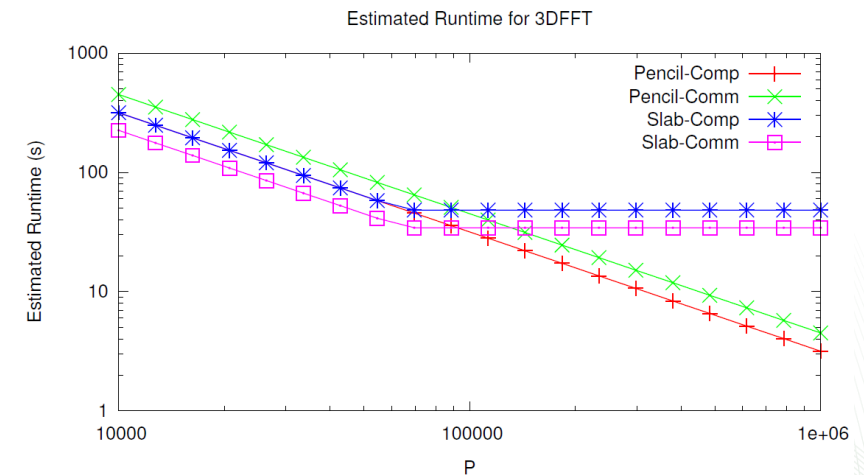
	Speed	Ease	Flexibility	Accuracy	Scalability
Ad-hoc Analytical Models	1	3	2	4	1
Structured Analytical Models	1	2	1	4	1
<i>Aspen</i>	1	1	1	4	1
Simulation – Functional	3	2	2	3	3
Simulation – Cycle Accurate	4	2	2	2	4
Hardware Emulation (FPGA)	3	3	3	2	3
Similar hardware measurement	2	1	4	2	2
Node Prototype	2	1	4	1	4
Prototype at Scale	2	1	4	1	2
Final System	-	-	-	-	-

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

Aspen Design Flow

Source code

```

2324 static inline
2325 void CalcMonotonicQGradientsForElems(Index_t p_nodelist[T_NUMELEM8],
2326   Real_t p_x[T_NUMNODE], Real_t p_y[T_NUMNODE], Real_t p_z[T_NUMNODE],
2327   Real_t p_xd[T_NUMNODE], Real_t p_yd[T_NUMNODE], Real_t p_zd[T_NUMNODE],
2328   Real_t p_volo[T_NUMELEM], Real_t p_vnew[T_NUMELEM],
2329   Real_t p_delx_zeta[T_NUMELEM], Real_t p_delv_zeta[T_NUMELEM],
2330   Real_t p_delx_xi[T_NUMELEM], Real_t p_delv_xi[T_NUMELEM],
2331   Real_t p_delx_eta[T_NUMELEM], Real_t p_delv_eta[T_NUMELEM])
2332 {
2333   Index_t i;
2334   Index_t numElem = m_numElem;
2335   #pragma acc parallel loop independent present(p_vnew, p_nodelist, p_x, p_y, p_z, p_xd, \
2336     p_yd, p_zd, p_volo, p_delx_xi, p_delx_eta, p_delx_zeta, p_delv_xi, p_delv_eta, \
2337     p_delv_zeta)
2338   for (i = 0 ; i < numElem ; ++i) {
2339     const Real_t ptiny = 1.e-36 ;
2340     Real_t ax, ay, az ;
2341     Real_t dxv, dyv, dzv ;
2342
2343     const Index_t *elemToNode = &p_nodelist[8*i];
2344     Index_t n0 = elemToNode[0] ;
2345     Index_t n1 = elemToNode[1] ;
2346     Index_t n2 = elemToNode[2] ;
2347     Index_t n3 = elemToNode[3] ;
2348     Index_t n4 = elemToNode[4] ;
2349     Index_t n5 = elemToNode[5] ;
2350     Index_t n6 = elemToNode[6] ;
2351     Index_t n7 = elemToNode[7] ;
2352
2353     Real_t x0 = p_x[n0] ;

```

Creation

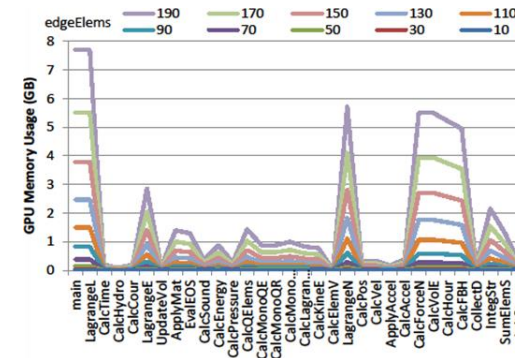
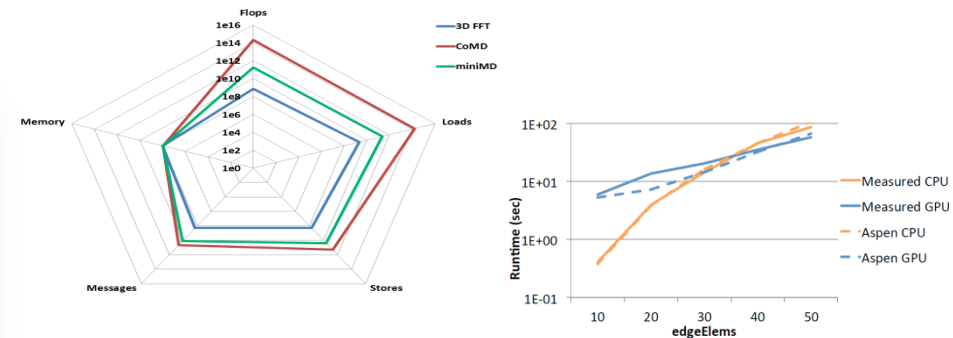
- Manual for future applications
- Static analysis via compilers
- Historical
- Empirical

Aspen code

```

147 kernel CalcMonotonicQGradients {
148     execute [numElems]
149     {
150         loads [8 * indexWordSize] from nodelist
151         // Load and cache position and velocity.
152         loads/caching [8 * wordSize] from x
153         loads/caching [8 * wordSize] from yvel
154         loads/caching [8 * wordSize] from z
155
156         loads/caching [8 * wordSize] from xvel
157         loads/caching [8 * wordSize] from yvel
158         loads/caching [8 * wordSize] from zvel
159
160         loads [wordSize] from volo
161         loads [wordSize] from vnew
162         // dx, dy, etc.
163         flops [90] as dp, simd
164         // delvk delkx
165         flops [9 + 8 + 3 + 30 + 5] as dp, simd
166         stores [wordSize] to delv_xeta
167         // delxi delvi
168         flops [9 + 8 + 3 + 30 + 5] as dp, simd
169         stores [wordSize] to delx_xi
170         // delxj and delvj
171         flops [9 + 8 + 3 + 30 + 5] as dp, simd
172         stores [wordSize] to delv_eta
173     }
174 }

```



Use

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

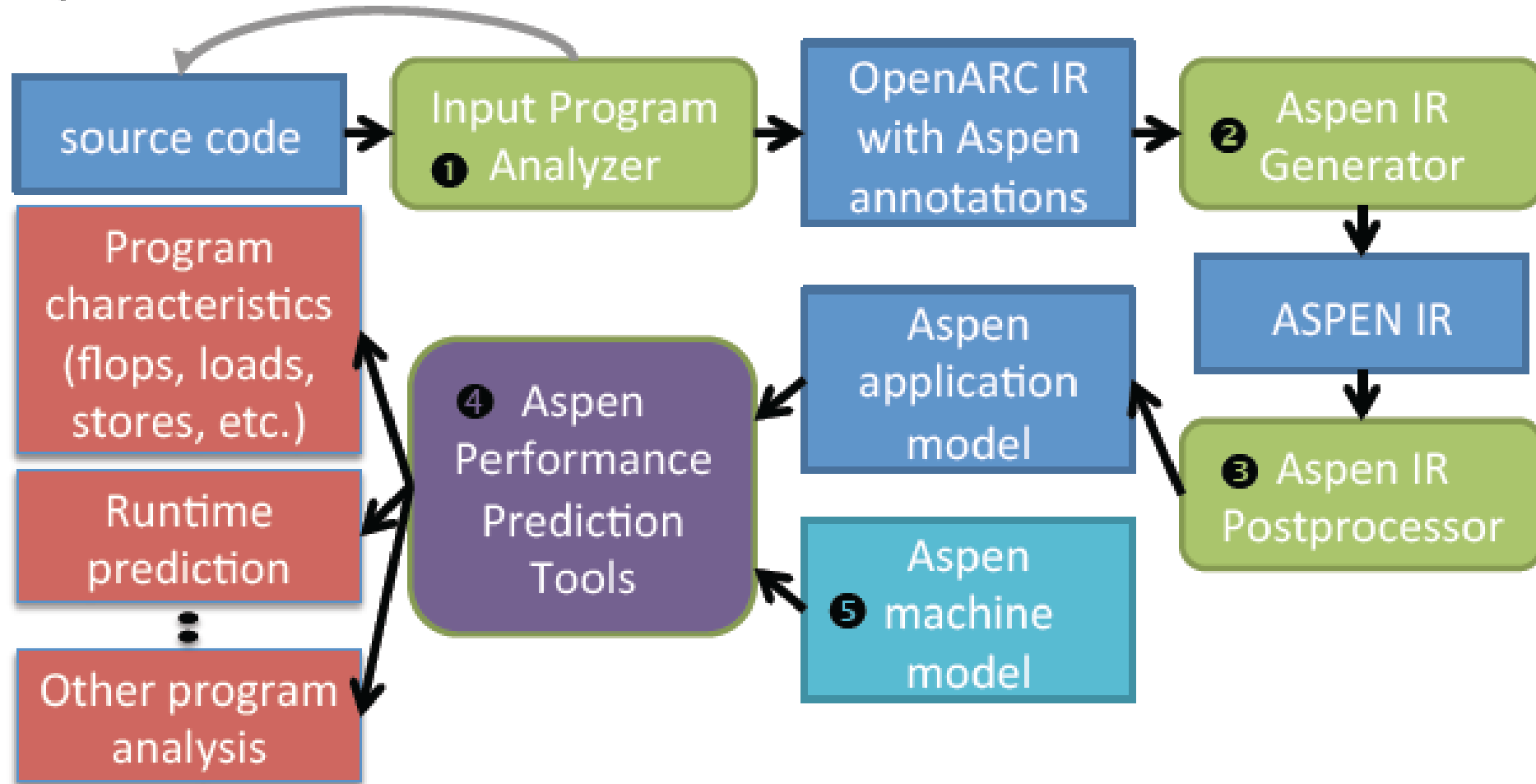
Representation in Aspen

- **Modular**
- **Sharable**
- **Composable**
- **Reflects prog structure**

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

Creating Aspen Models

Optional feedback for advanced users




Simple MM example generated from COMPASS

```
1 int N = 1024;
2 void matmul(float *a, float *b, float *c){ int i, j, k ;
3 #pragma acc kernels loop gang copyout(a[0:(N*N)]) \
4 copyin(b[0:(N*N)],c[0:(N*N)])
5   for (i=0; i<N; i++){
6   #pragma acc loop worker
7     for (j=0; j<N; j++) { float sum = 0.0 ;
8       for (k=0; k<N; k++) {sum+=b[i*N+k]*c[k*N+j];}
9       a[i*N+j] = sum; }
10  } //end of i loop
11 } //end of matmul()
12 int main() {
13   int i; float *A = (float*) malloc(N*N*sizeof(float));
14   float *B = (float*) malloc(N*N*sizeof(float));
15   float *C = (float*) malloc(N*N*sizeof(float));
16   for (i = 0; i < N*N; i++)
17     { A[i] = 0.0F; B[i] = (float) i; C[i] = 1.0F; }
18   #pragma aspen modelregion label(MM)
19   matmul(A,B,C);
20   free(A); free(B); free(C); return 0;
21 } //end of main()
```

```
1 model MM {
2   param floatS = 4; param N = 1024
3   data A as Array((N*N), floatS)
4   data B as Array((N*N), floatS)
5   data C as Array((N*N), floatS)
6   kernel matmul {
7     execute matmul2_intracommIN
8     { intracomm [floatS*(N*N)] to C as copyin
9       intracomm [floatS*(N*N)] to B as copyin }
10  map matmul2 [N] {
11    map matmul3 [N] {
12      iterate [N] {
13        execute matmul5
14        { loads [floatS] from B as stride(1)
15          loads [floatS] from C; flops [2] as sp, simd }
16      } //end of iterate
17      execute matmul6 { stores [floatS] to A as stride(1) }
18    } // end of map matmul3
19  } //end of map matmul2
20  execute matmul2_intracommOUT
21  { intracomm [floatS*(N*N)] to A as copyout }
22 } //end of kernel matmul
23 kernel main { matmul() }
24 } //end of model MM
```


LULESH in Aspen

branch: master **aspen / models / lulesh / lulesh.aspen**

 **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)
```

```
147 kernel CalcMonotonicQGradients {
148     execute [numElems]
149     {
150         loads [8 * indexWordSize] from nodelist
151         // Load and cache position and velocity.
152         loads/caching [8 * wordSize] from x
153         loads/caching [8 * wordSize] from y
154         loads/caching [8 * wordSize] from z
155
156         loads/caching [8 * wordSize] from xvel
157         loads/caching [8 * wordSize] from yvel
158         loads/caching [8 * wordSize] from zvel
159
160         loads [wordSize] from volo
161         loads [wordSize] from vnew
162         // dx, dy, etc.
163         flops [90] as dp, simd
164         // delvk delxk
165         flops [9 + 8 + 3 + 30 + 5] as dp, simd
166         stores [wordSize] to delv_xeta
167         // delxi delvi
168         flops [9 + 8 + 3 + 30 + 5] as dp, simd
169         stores [wordSize] to delx_xi
170         // delxj and delvj
171         flops [9 + 8 + 3 + 30 + 5] as dp, simd
172         stores [wordSize] to delv_eta
173     }
174 }
```

LULESH – runtime optimizations

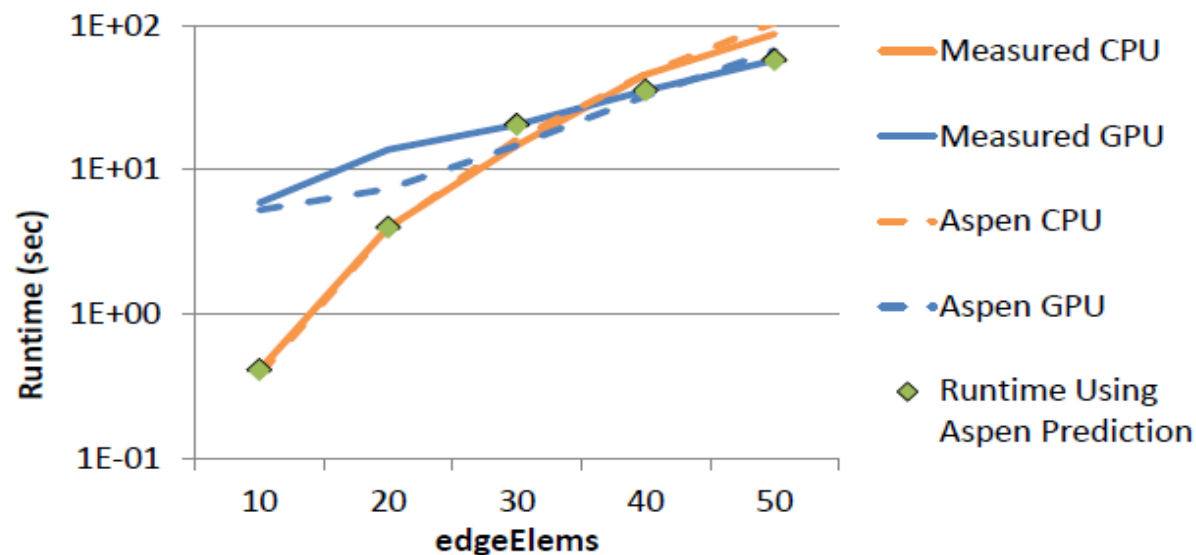


Fig. 7: Measured and predicted runtime of the entire LULESH program on CPU and GPU, including measured runtimes using the automatically predicted optimal target device at each size.

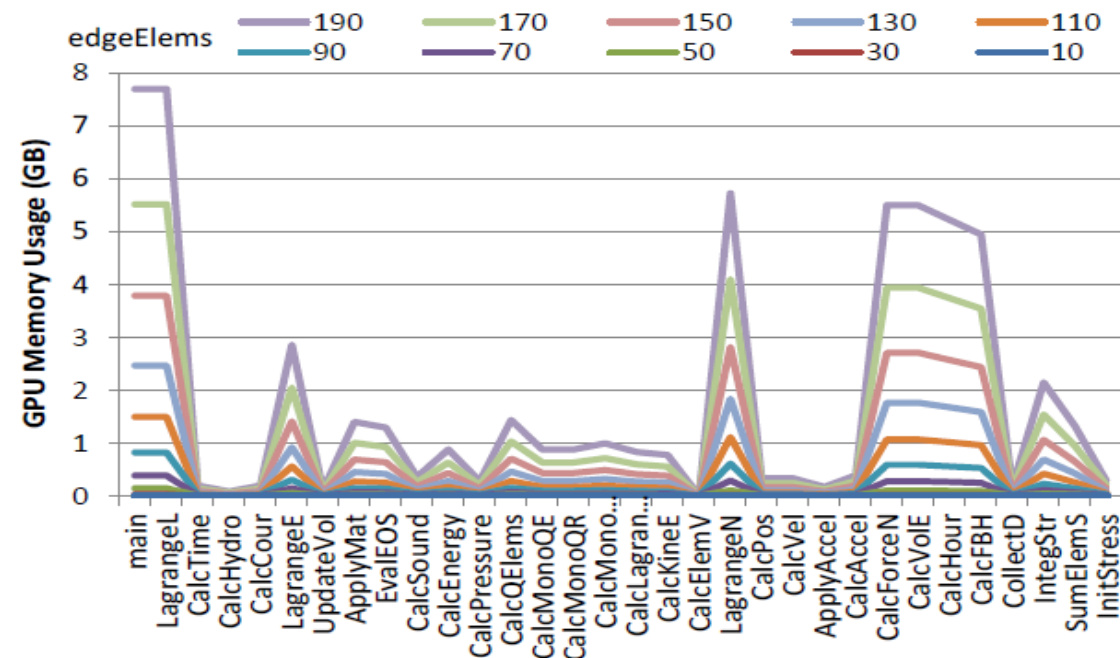


Fig. 8: GPU Memory Usage of each Function in LULESH, where the memory usage of a function is inclusive; value for a parent function includes data accessed by its child functions in the call graph.

3DFFT

```
// Dimension of cubic 3D Volume
param n = 8192
param a = 6.3
param wordSize = 16 // Double Complex Words
param dataPerProc = (n^3 * wordSize) / P
data fftVolume [n^3 * wordSize]
```

```
control pencil {
  localFFT -> transpose // in X
  exchange
  localFFT -> transpose // in Y
  exchange
  localFFT -> transpose // in Z
}
```

```
control slab {
  localFFT -> transpose // in X
  localFFT -> transpose // in Y
  exchange
  localFFT -> transpose // in Z
}
```

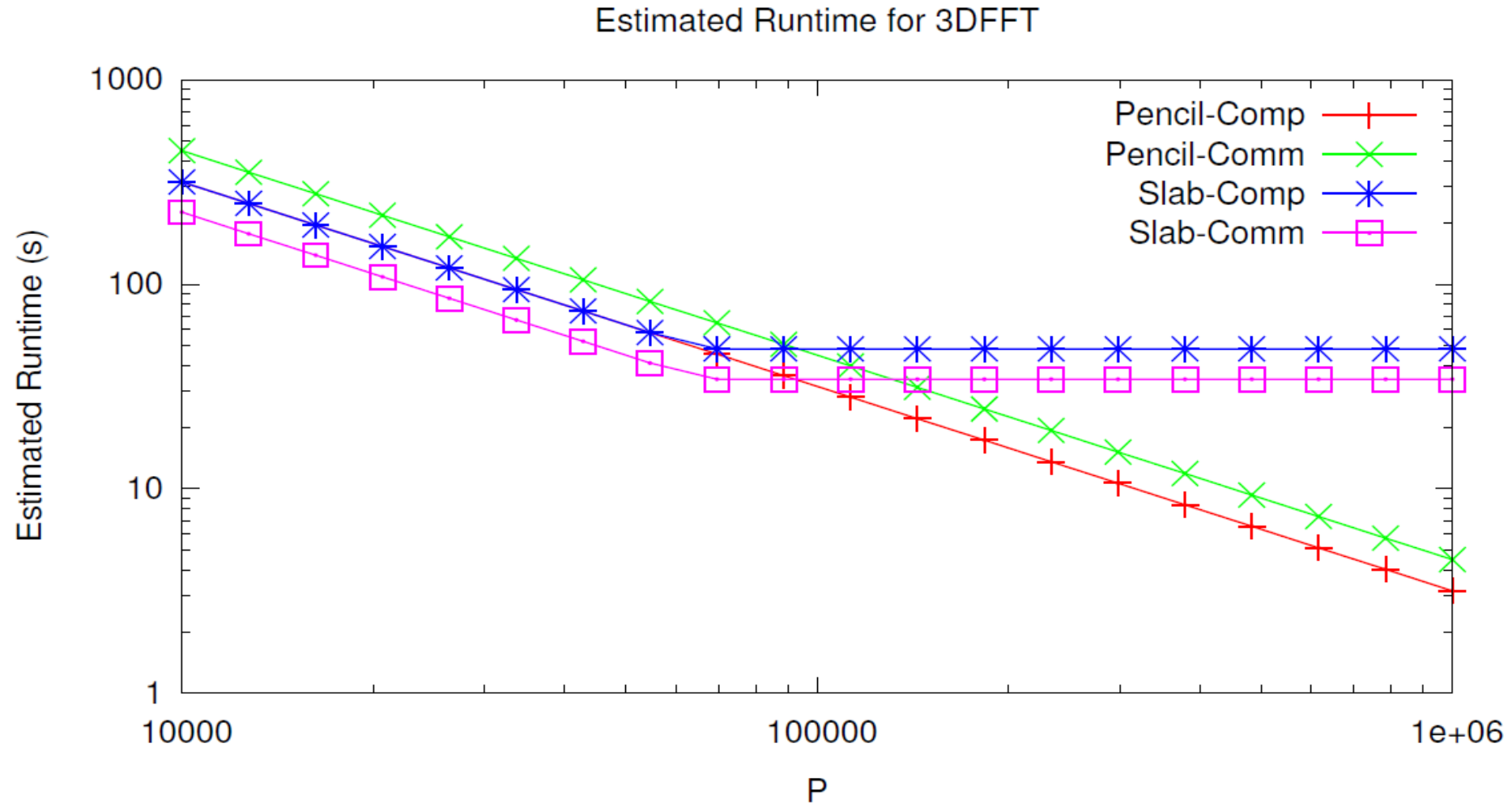
```
kernel localFFT {
  exposes parallelism [n^2]
  requires flops [5 * n * log2(n)] as dp, simd
  requires loads [a * (n*wordSize) * max(1, log(n*wordSize)/log(Z))]
    from fftVolume
}
```

```
kernel exchange {
  exposes parallelism [P]
  requires messages [(n^3 * wordSize) / P] as
    allToAll
}
```

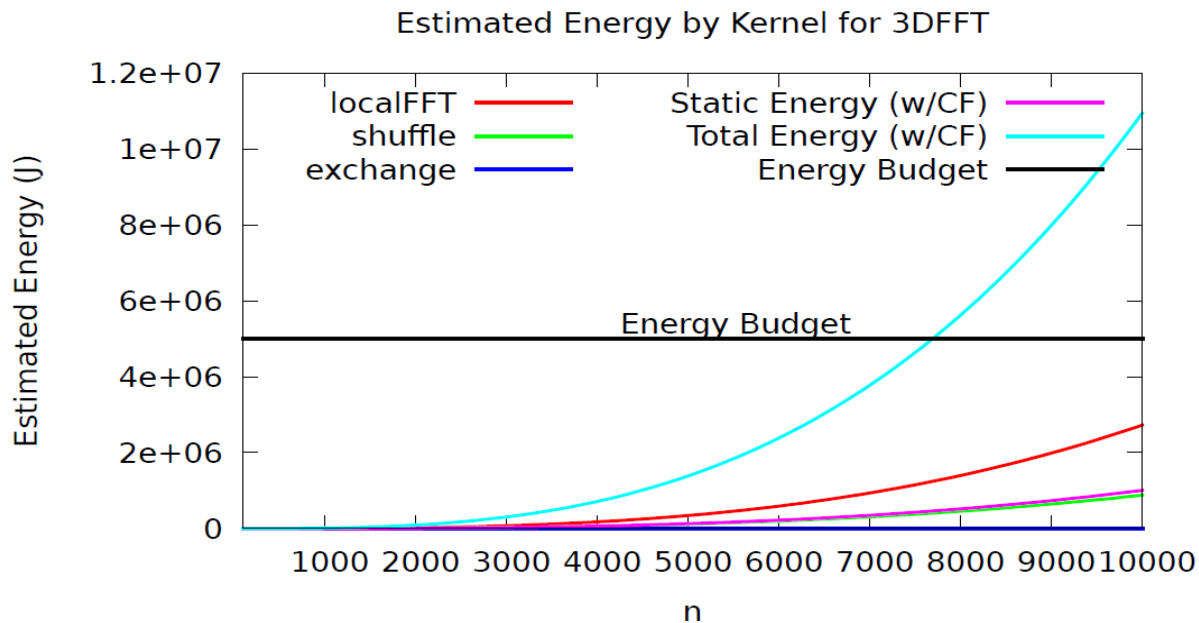
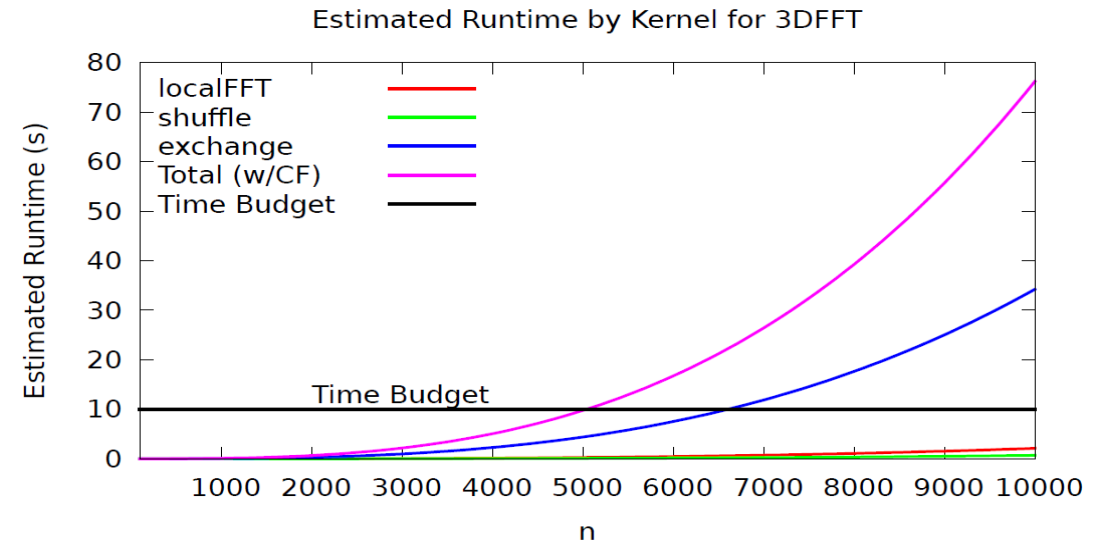
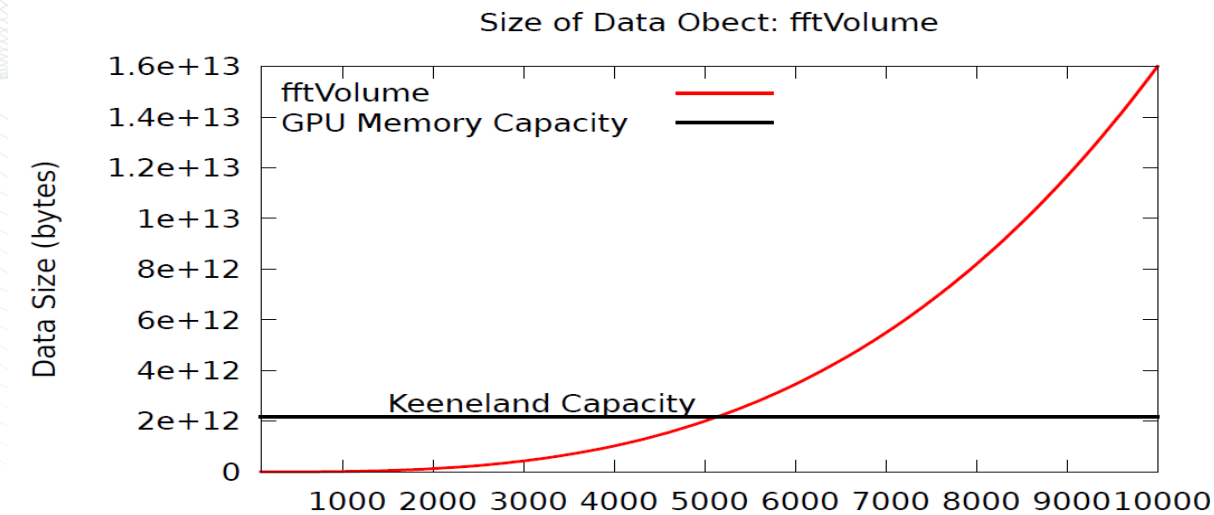
3DFFT: Slab vs. Pencil Tradeoff

Ideal Parallelism

- Insights become obvious with Aspen

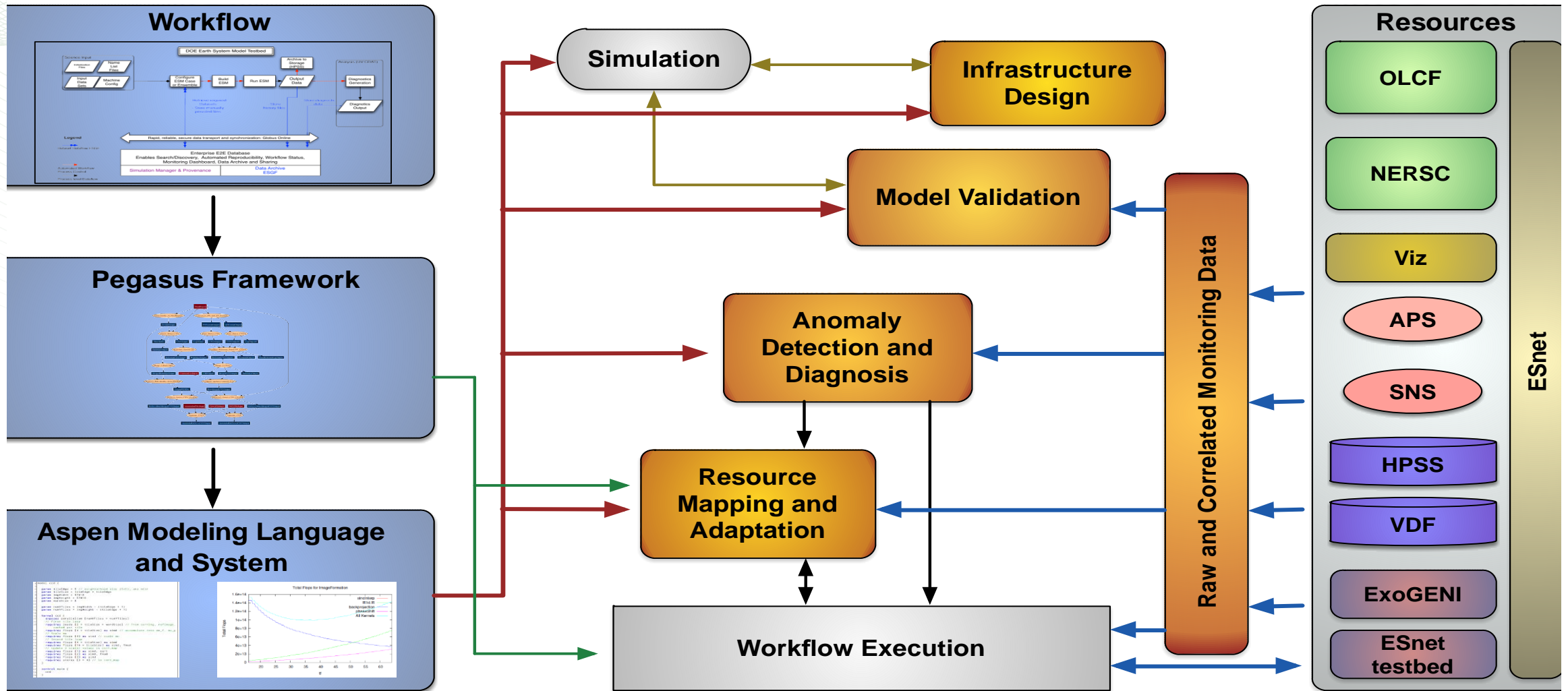


Design Space Exploration

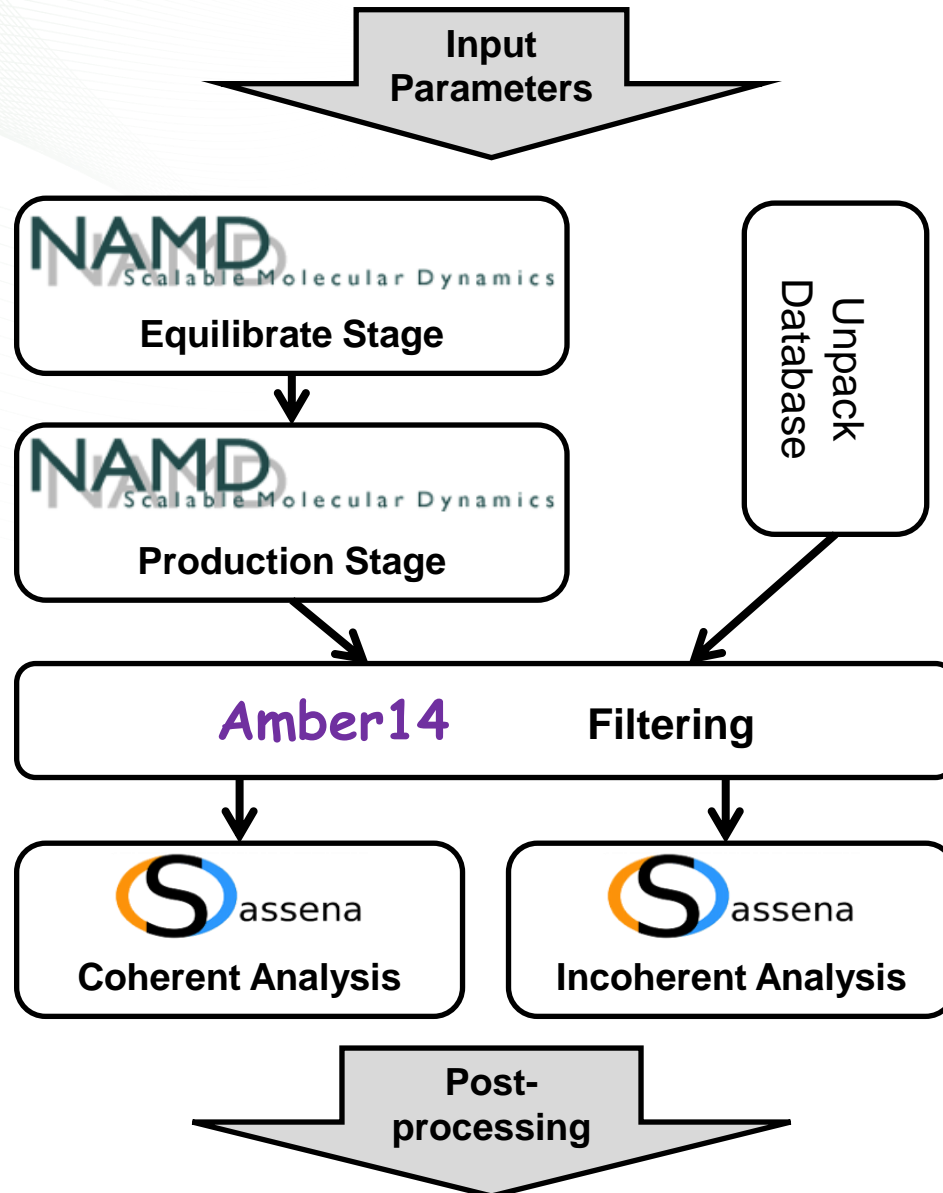


- n is approximately 5000

PANORAMA Overview



Spallation Neutron Source Workflow



```
kernel main
{
    par {
        seq {
            call namd_eq_200()
            call namd_prod_200()
        }
        seq {
            call namd_eq_290()
            call namd_prod_290()
        }
        call unpack_database()
    }
    par {
        call amber_ptraj_200()
        call amber_ptraj_290()
    }
    par {
        call sassena_coh_200()
        call sassena_coh_290()
        call sassena_inc_200()
        call sassena_inc_290()
    }
}
```

Summary

- Our community has major challenges in HPC as we move to extreme scale
 - Power, Performance, Resilience, Productivity
 - Major shifts in architectures, software, applications
 - Not just HPC: Most uncertainty in two decades
- New technologies emerging to address some of these challenges
 - Heterogeneous computing
 - Nonvolatile memory
- Consequently, we now have critical situations in
 - Portable programming models
 - Performance prediction for procurement, optimization, etc
- Aspen is a tool we have developed for performance prediction

Acknowledgements

- Contributors and Sponsors
 - Future Technologies Group: <http://ft.ornl.gov>
 - US Department of Energy Office of Science
 - DOE Vancouver Project: <https://ft.ornl.gov/trac/vancouver>
 - DOE Blackcomb Project: <https://ft.ornl.gov/trac/blackcomb>
 - DOE ExMatEx Codesign Center: <http://codesign.lanl.gov>
 - DOE Cesar Codesign Center: <http://cesar.mcs.anl.gov/>
 - DOE Exascale Efforts: <http://science.energy.gov/ascr/research/computer-science/>
 - Scalable Heterogeneous Computing Benchmark team: <http://bit.ly/shocmarx>
 - US National Science Foundation Keeneland Project: <http://keeneland.gatech.edu>
 - US DARPA
 - NVIDIA CUDA Center of Excellence

