SIMULATION OF EXASCALE NODES THROUGH RUNTIME HARDWARE MONITORING

JOSEPH L. GREATHOUSE, ALEXANDER LYASHEVSKY, MITESH MESWANI, NUWAN JAYASENA, MICHAEL IGNATOWSKI

9/18/2013

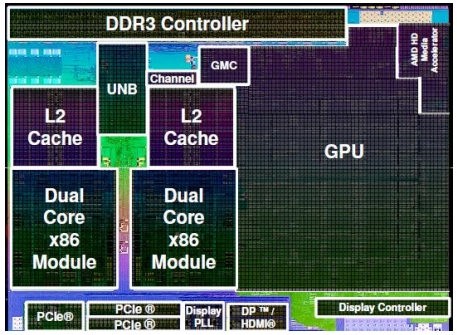
CHALLENGES OF EXASCALE NODE RESEARCH

MANY DESIGN DECISIONS

CHALLENGES OF EXASCALE NODE RESEARCH

MANY DESIGN DECISIONS

Heterogeneous Cores

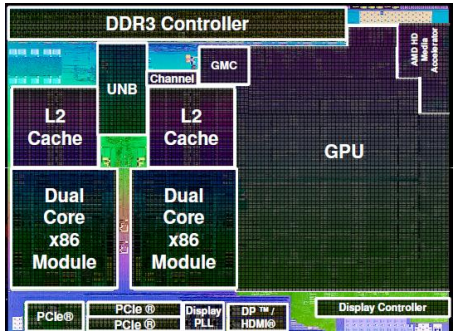


Composition? Size? Speed?

CHALLENGES OF EXASCALE NODE RESEARCH

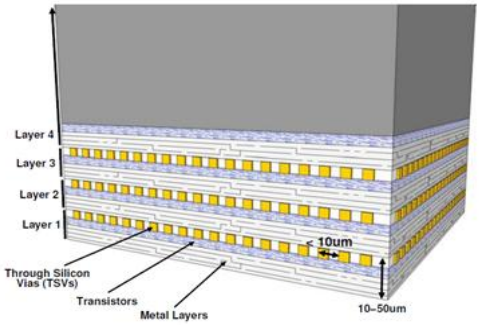
MANY DESIGN DECISIONS

Heterogeneous Cores



Composition? Size? Speed?

Stacked Memories

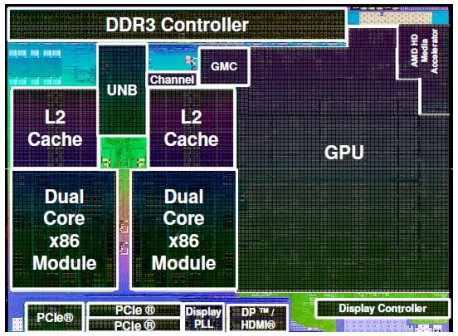


Useful? Compute/BW Ratio? Latency? Capacity? Non-Volatile?

CHALLENGES OF EXASCALE NODE RESEARCH

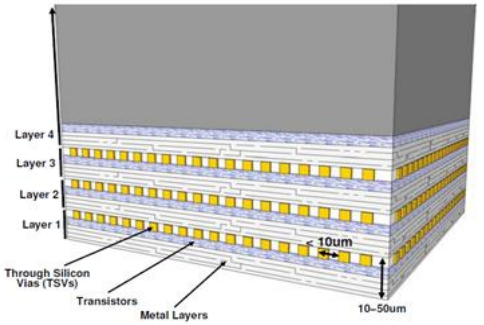
MANY DESIGN DECISIONS

Heterogeneous Cores



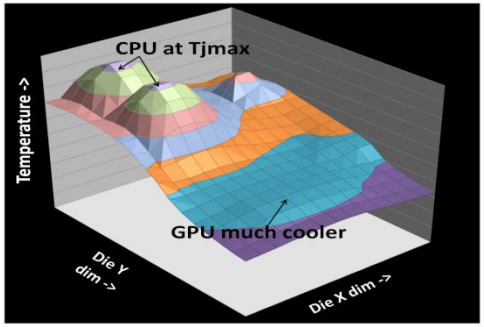
Composition? Size? Speed?

Stacked Memories



Useful? Compute/BW Ratio? Latency? Capacity? Non-Volatile?

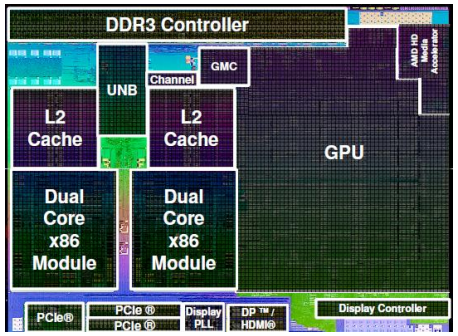
Thermal Constraints



CHALLENGES OF EXASCALE NODE RESEARCH

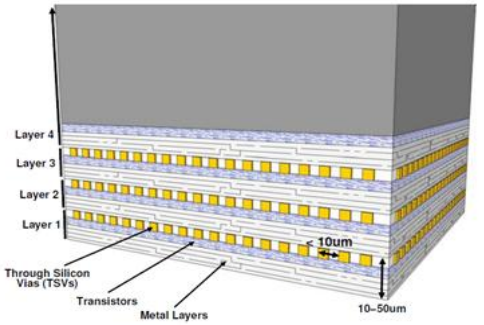
MANY DESIGN DECISIONS

Heterogeneous Cores



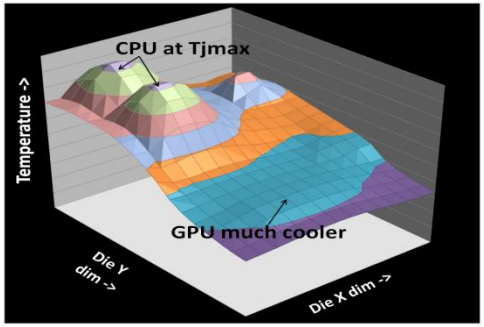
Composition? Size? Speed?

Stacked Memories



Useful? Compute/BW Ratio? Latency? Capacity? Non-Volatile?

Thermal Constraints



Power Sharing? Heat dissipation? Sprinting?

Software Co-Design

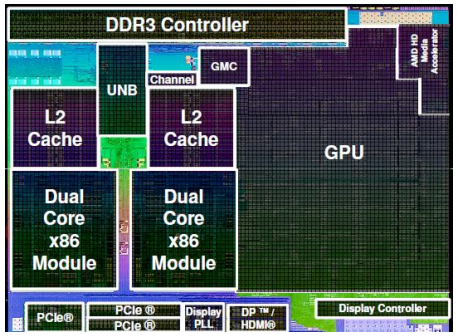
```
Real_t vol = volo[i]*vnew[i] ;
Real_t norm = (Real_t)(1.0) / ( vol + ptiny ) ;

Real_t dxj = (Real_t)(-0.25)*(SUM4(x0,x1,x5,x4) - SUM4(x3,x2,x6,x7)) ;
Real_t dyj = (Real_t)(-0.25)*(SUM4(y0,y1,y5,y4) - SUM4(y3,y2,y6,y7)) ;
Real_t dzj = (Real_t)(-0.25)*(SUM4(z0,z1,z5,z4) - SUM4(z3,z2,z6,z7)) ;

Real_t dxi = (Real_t)( 0.25)*(SUM4(x1,x2,x6,x5) - SUM4(x0,x3,x7,x4)) ;
Real_t dyi = (Real_t)( 0.25)*(SUM4(y1,y2,y6,y5) - SUM4(y0,y3,y7,y4)) ;
Real_t dzi = (Real_t)( 0.25)*(SUM4(z1,z2,z6,z5) - SUM4(z0,z3,z7,z4)) ;

Real_t dxk = (Real_t)( 0.25)*(SUM4(x4,x5,x6,x7) - SUM4(x0,x1,x2,x3)) ;
Real_t dyk = (Real_t)( 0.25)*(SUM4(y4,y5,y6,y7) - SUM4(y0,y1,y2,y3)) ;
Real_t dzk = (Real_t)( 0.25)*(SUM4(z4,z5,z6,z7) - SUM4(z0,z1,z2,z3)) ;
```

New algorithms? Data placement? Programming models?

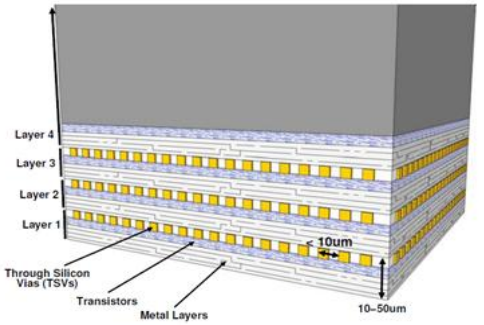
CHALLENGES OF EXASCALE NODE RESEARCH

MANY DESIGN DECISIONS

Heterogeneous Cores

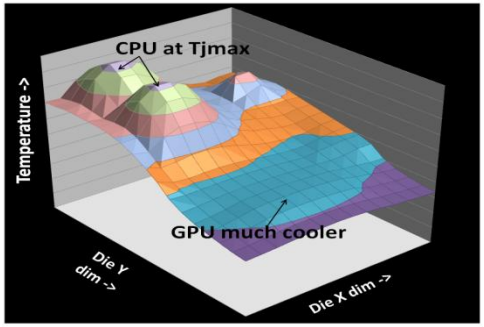


Stacked Memories



Com

Exascale: Huge Design Space to Explore



Power Sharing? Heat dissipation? Sprinting?

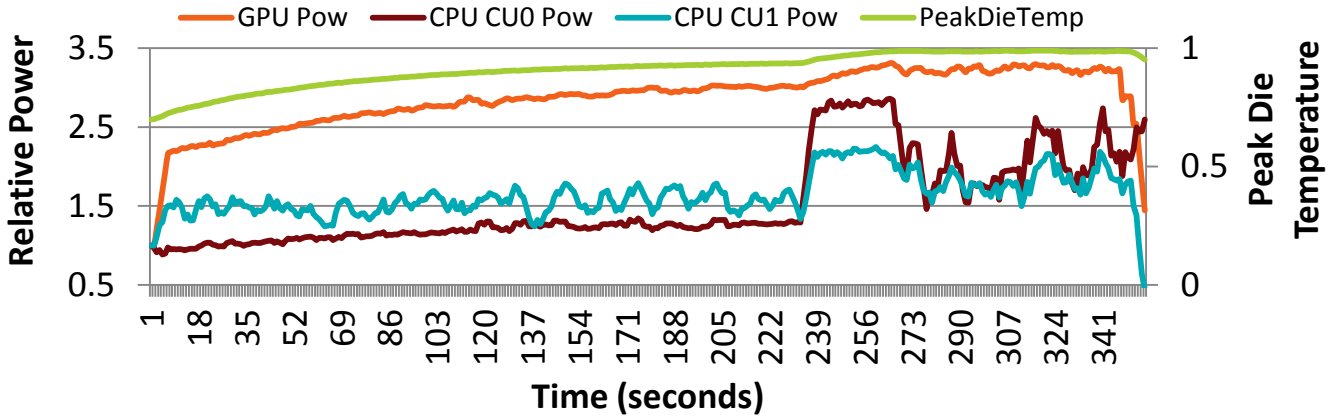
```
Real_t vol = volo[i]*vnew[i] ;
Real_t norm = (Real_t)(1.0) / ( vol + ptiny ) ;

Real_t dxj = (Real_t)(-0.25)*(SUM4(x0,x1,x5,x4) - SUM4(x3,x2,x6,x7)) ;
Real_t dyj = (Real_t)(-0.25)*(SUM4(y0,y1,y5,y4) - SUM4(y3,y2,y6,y7)) ;
Real_t dzj = (Real_t)(-0.25)*(SUM4(z0,z1,z5,z4) - SUM4(z3,z2,z6,z7)) ;

Real_t dxi = (Real_t)( 0.25)*(SUM4(x1,x2,x6,x5) - SUM4(x0,x3,x7,x4)) ;
Real_t dyi = (Real_t)( 0.25)*(SUM4(y1,y2,y6,y5) - SUM4(y0,y3,y7,y4)) ;
Real_t dzi = (Real_t)( 0.25)*(SUM4(z1,z2,z6,z5) - SUM4(z0,z3,z7,z4)) ;

Real_t dxk = (Real_t)( 0.25)*(SUM4(x4,x5,x6,x7) - SUM4(x0,x1,x2,x3)) ;
Real_t dyk = (Real_t)( 0.25)*(SUM4(y4,y5,y6,y7) - SUM4(y0,y1,y2,y3)) ;
Real_t dzk = (Real_t)( 0.25)*(SUM4(z4,z5,z6,z7) - SUM4(z0,z1,z2,z3)) ;
```

New algorithms? Data placement? Programming models?

CHALLENGES OF EXASCALE NODE RESEARCH

INTERESTING QUESTION REQUIRE LONG RUNTIMES

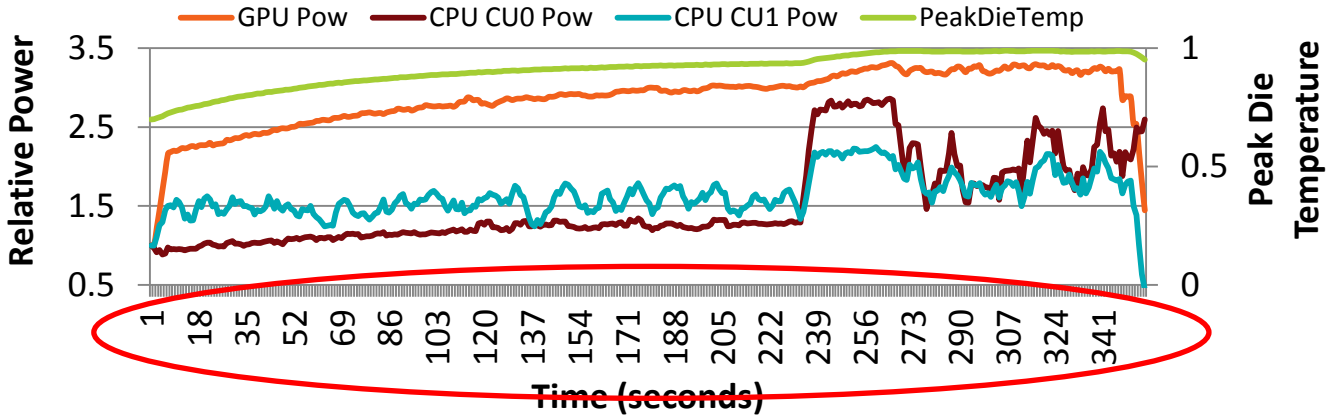
- Power and Thermals on a real heterogeneous processor:



CHALLENGES OF EXASCALE NODE RESEARCH

INTERESTING QUESTION REQUIRE LONG RUNTIMES

- Power and Thermals on a real heterogeneous processor:



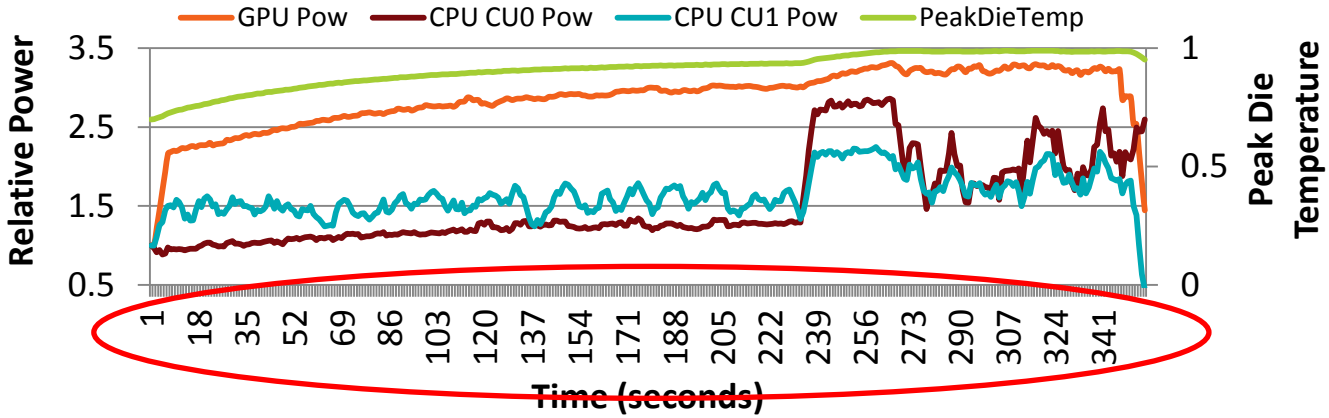
– ~2.5 trillion CPU instructions, ~60 trillion GPU operations

CHALLENGES OF EXASCALE NODE RESEARCH

INTERESTING QUESTION REQUIRE LONG RUNTIMES

AMD

- Power and Thermals on a real heterogeneous processor:



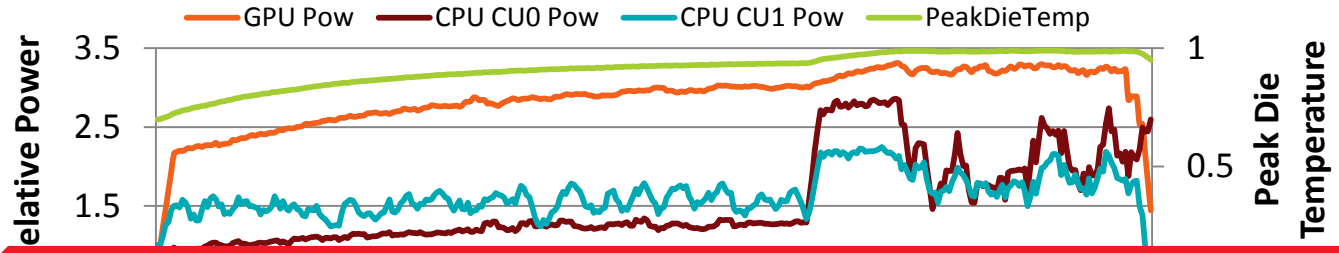
  - ~2.5 trillion CPU instructions, ~60 trillion GPU operations

- Exascale Proxy Applications are Large
  - Large initialization phases, many long iterations
  - Not microbenchmarks
  - **Already** reduced inputs and computation from real HPC applications

CHALLENGES OF EXASCALE NODE RESEARCH

INTERESTING QUESTION REQUIRE LONG RUNTIMES

- Power and Thermals on a real heterogeneous processor:



Exascale: Huge Execution Times

- Exascale Proxy Applications are Large
  - Large initialization phases, many long iterations
  - Not microbenchmarks
  - **Already** reduced inputs and computation from real HPC applications

EXISTING SIMULATORS

AMD

- Microarchitecture Sims: e.g. gem5, Multi2Sim, MARSSx86, SESC, GPGPU-Sim
  - Excellent for low-level details. We need these!
  - Too slow for design space explorations: **~60 trillion operations = 1 year of sim time**
- Functional Simulators: e.g. SimNow, Simics, QEMU, etc.
  - Faster than microarchitectural simulators, good for things like access patterns
  - No relation to hardware performance
- High-Level Simulators: e.g. Sniper, Graphite, CPR
  - Break operations down into timing models, e.g. core interactions, pipeline stalls, etc.
  - Faster, easier to parallelize.
  - Runtimes and complexity still constrained by desire to achieve accuracy.

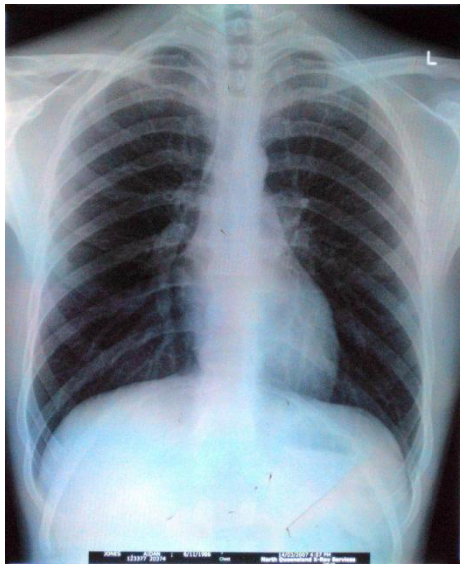
TRADE OFF INDIVIDUAL TEST ACCURACY

AMD

- Doctors do not start with:

http://www.flickr.com/photos/calliope/361953316/ - CC BY 2.0

http://www.flickr.com/photos/elsie/7080667207/ - CC BY 2.0

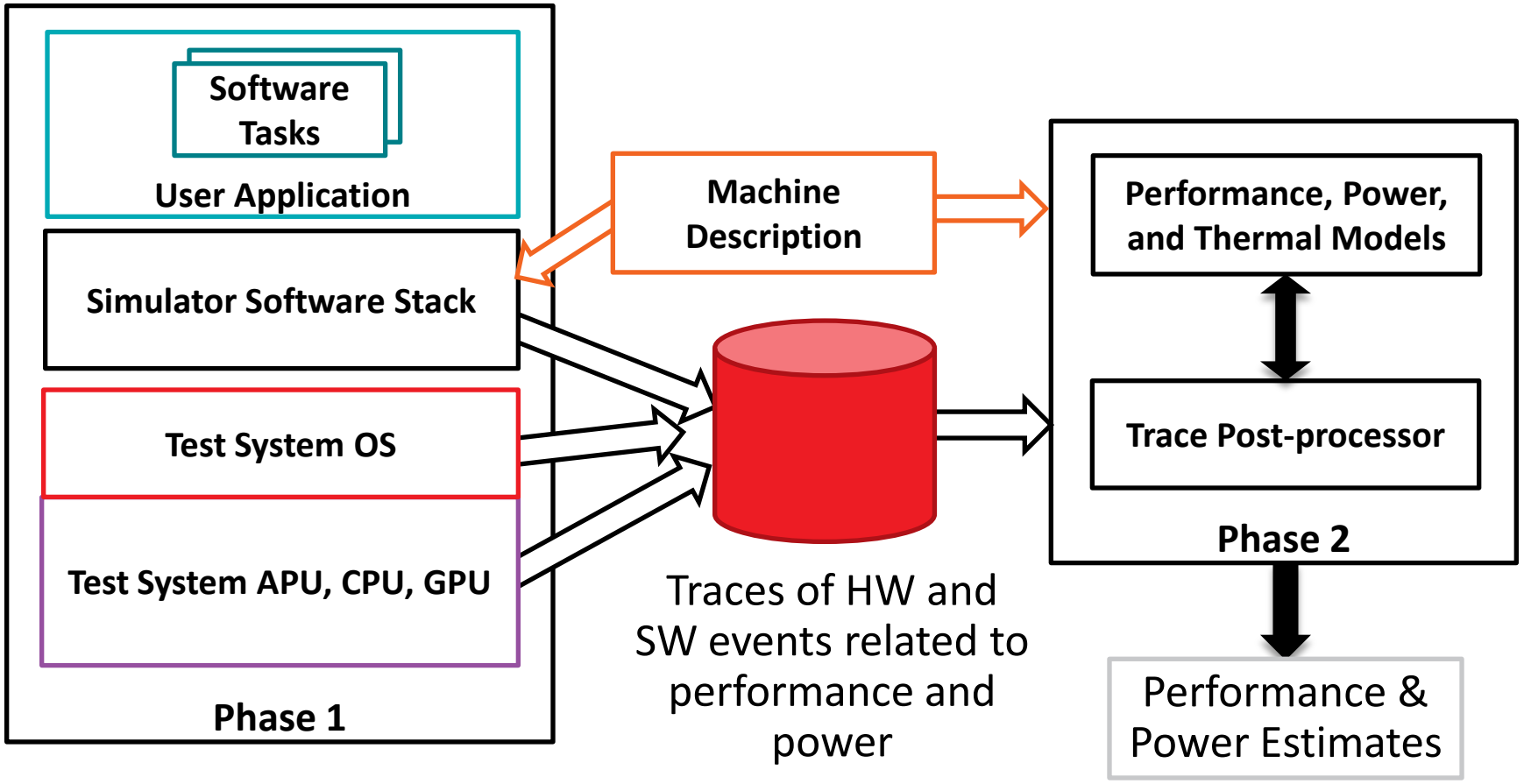


http://www.flickr.com/photos/aidan_jones/1438403889/
CC BY-SA 2.0

Fast Simulation Using Hardware Monitoring

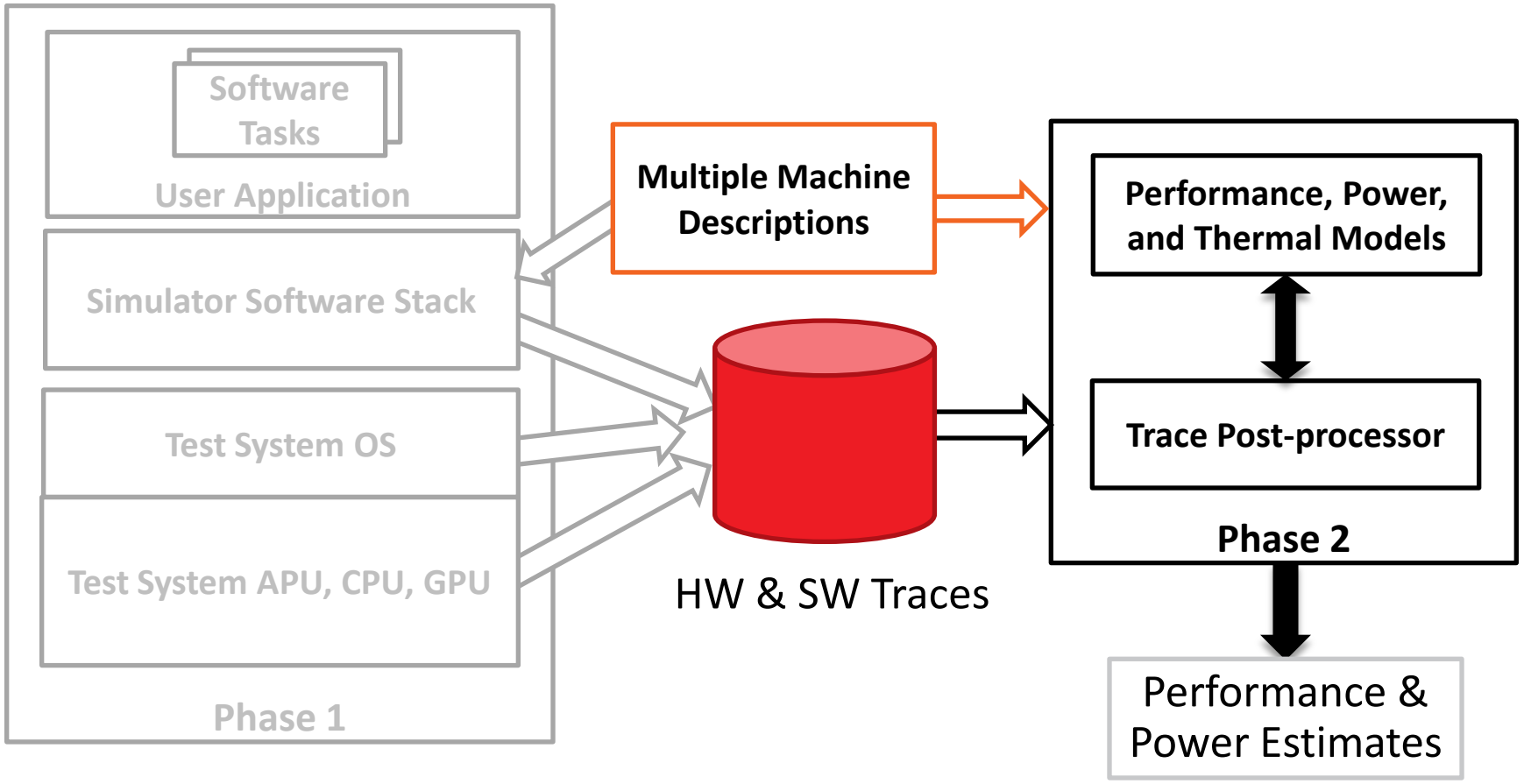
HIGH-LEVEL SIMULATION METHODOLOGY

MULTI-STAGE PERFORMANCE ESTIMATION PROCESS



BENEFITS OF MULTI-STEP PROCESS

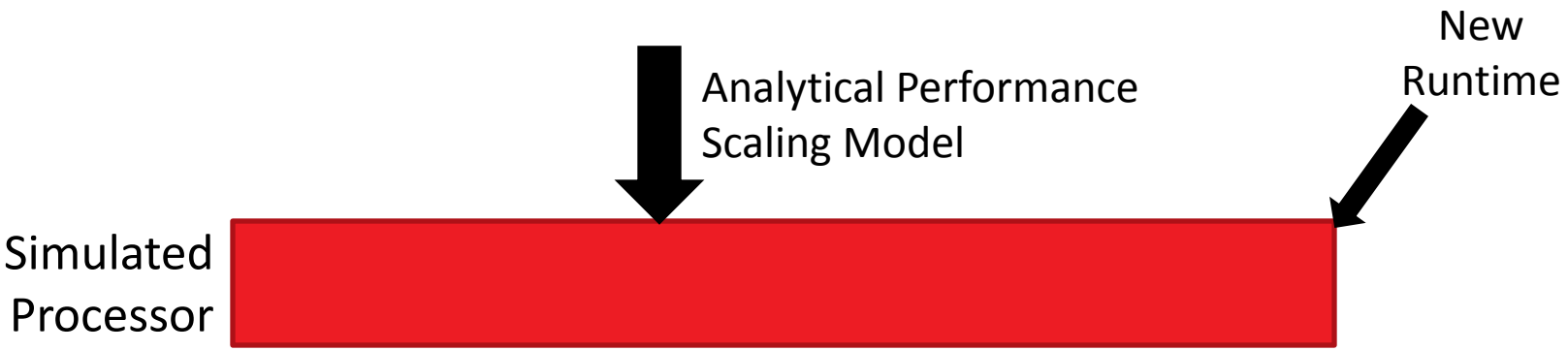
MANY DESIGN SPACE CHANGES ONLY NEED SECOND STEP



PERFORMANCE SCALING A SINGLE-THREADED APP

Gather statistics and performance counters about:

- Instructions Committed, stall cycles
- Memory operations, cache misses, etc.
- Power usage



ANALYTIC PERFORMANCE SCALING

AMD

- CPU Performance Scaling:
  - Find stall events using HW perf. counters. Scale based on new machine parameters.
  - Large amount of work in the literature on DVFS performance scaling, interval models..
  - Some events can be scaled by fiat:
    "If IPC when not stalled on memory doubled, what would happen to performance?"

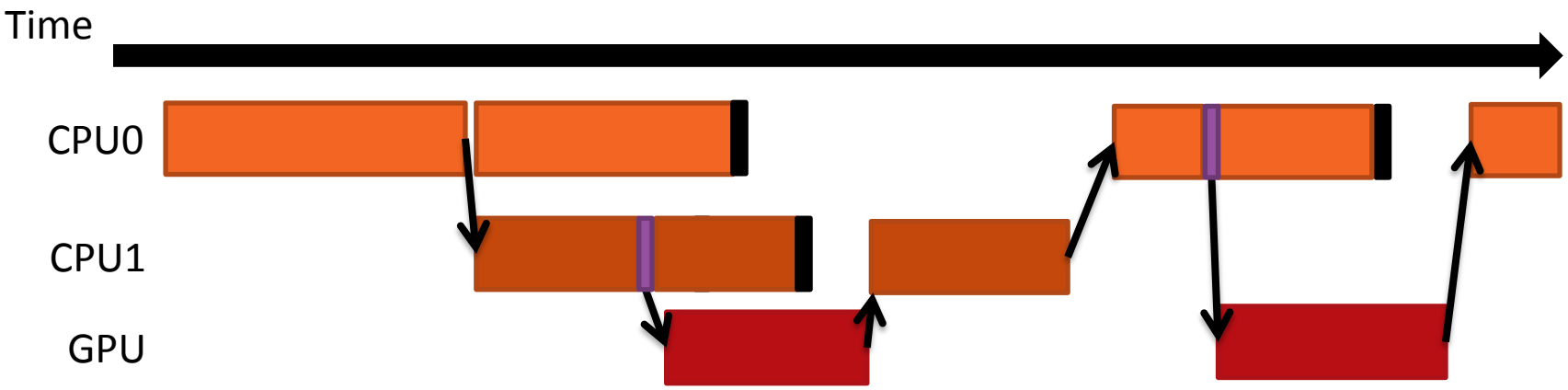
- GPU Performance Scaling:
  - Watch HW perf. counters that indicate work to do, memory usage, and GPU efficiency
  - Scale values based on estimations of parameters to test

OTHER ANALYTIC MODELS

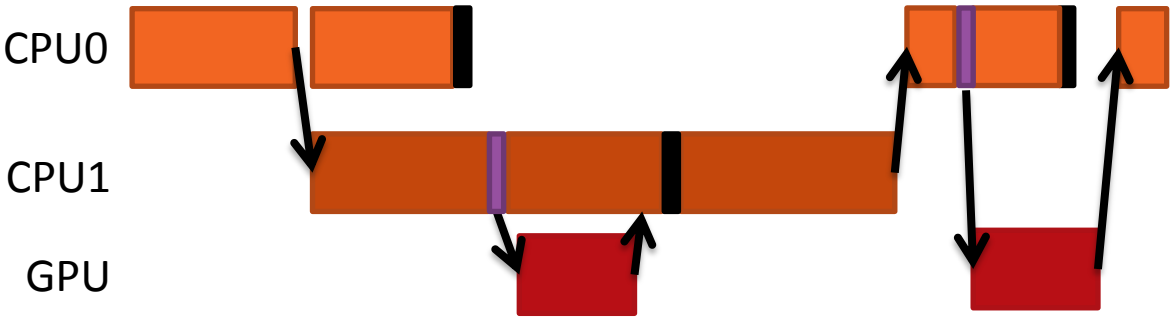
AMD

- Cache/Memory Access Model
  - Observe memory access traces using binary instrumentation or hardware
  - Send traces through high-level cache simulation system
  - Find knees in the curve, feed this back to CPU/GPU performance scaling model
- Power and thermal models
  - Power can be correlated to hardware events or directly measured
  - Scaled to future technology points
  - Any number of thermal models will work at this point
- Thermal and power models can feed into control algorithms that change system performance
  - This is **another** HW/SW co-design point. Fast operation is essential.

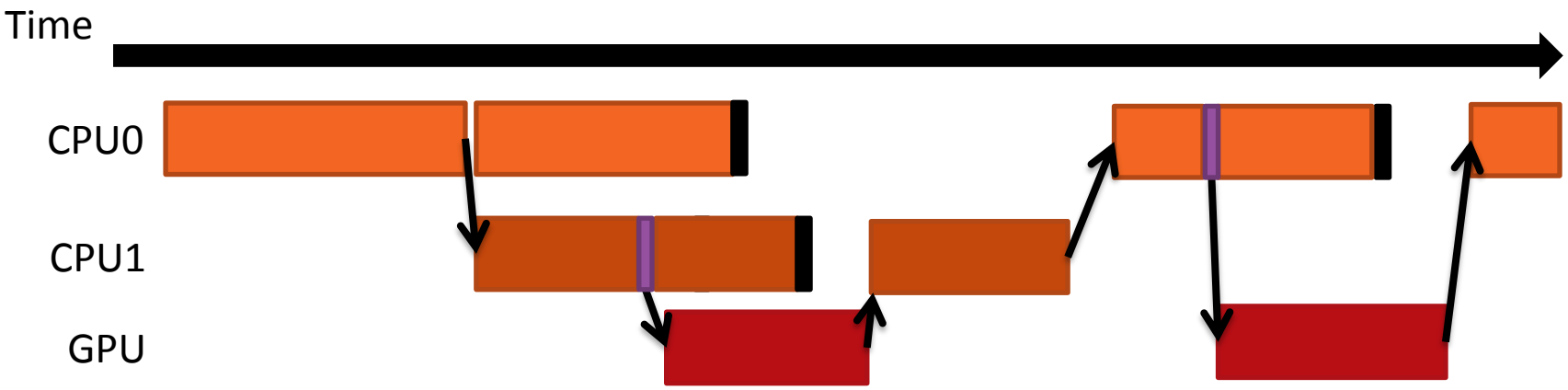
SCALING PARALLEL SEGMENTS REQUIRES MORE INFO



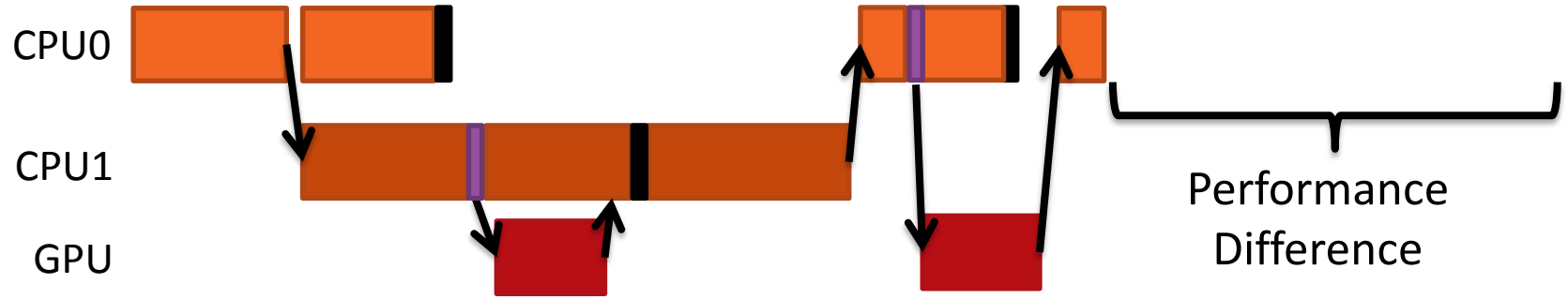
Example when everything except CPU1 gets faster:



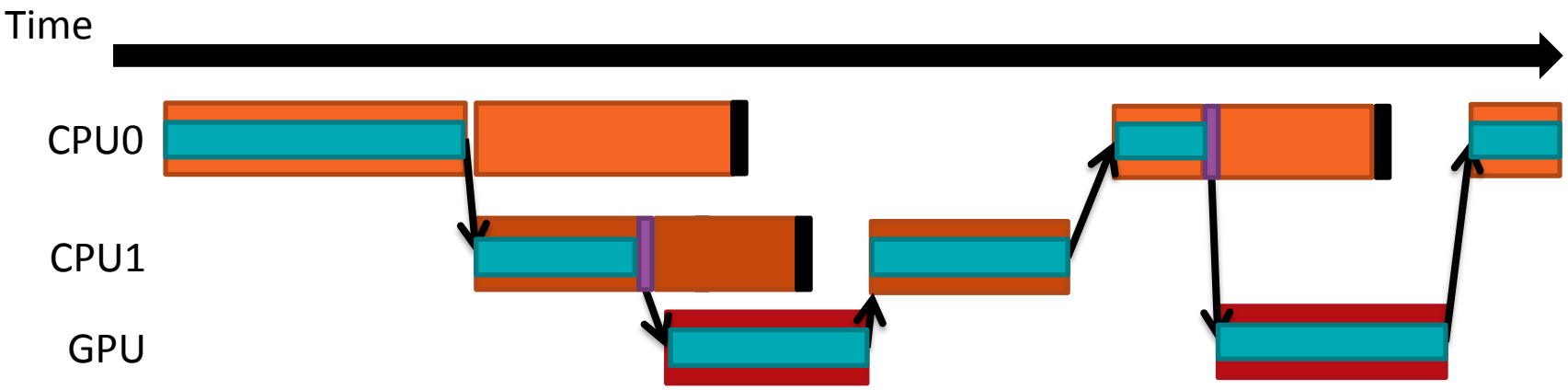
SCALING PARALLEL SEGMENTS REQUIRES MORE INFO



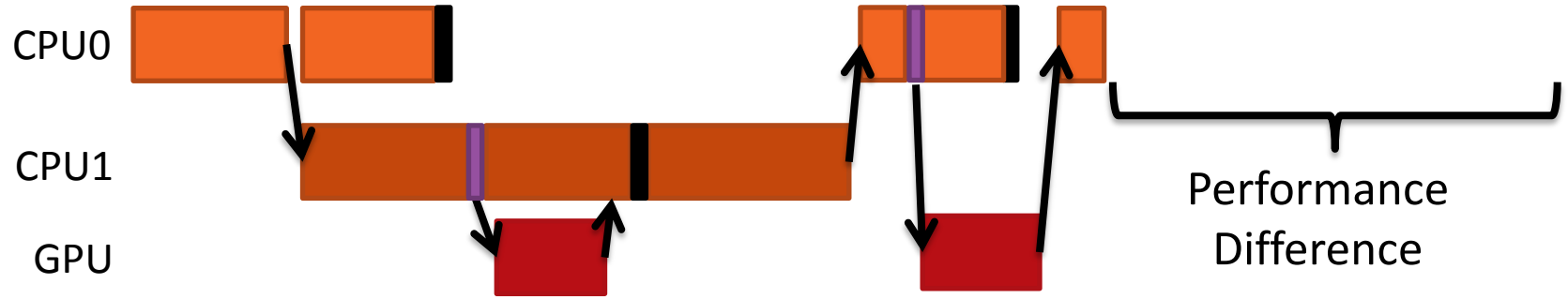
Example when everything except CPU1 gets faster:



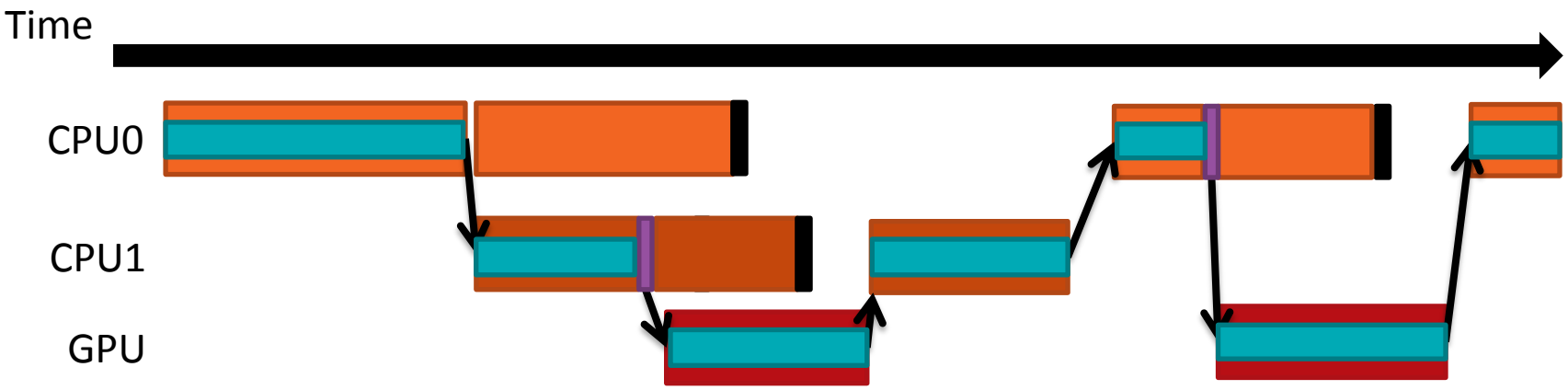
SCALING PARALLEL SEGMENTS REQUIRES MORE INFO



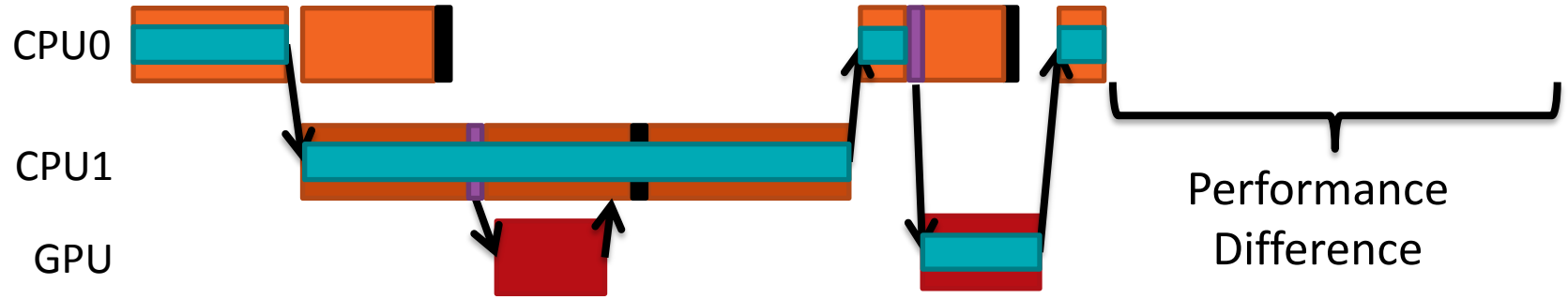
Example when everything except CPU1 gets faster:



SCALING PARALLEL SEGMENTS REQUIRES MORE INFO

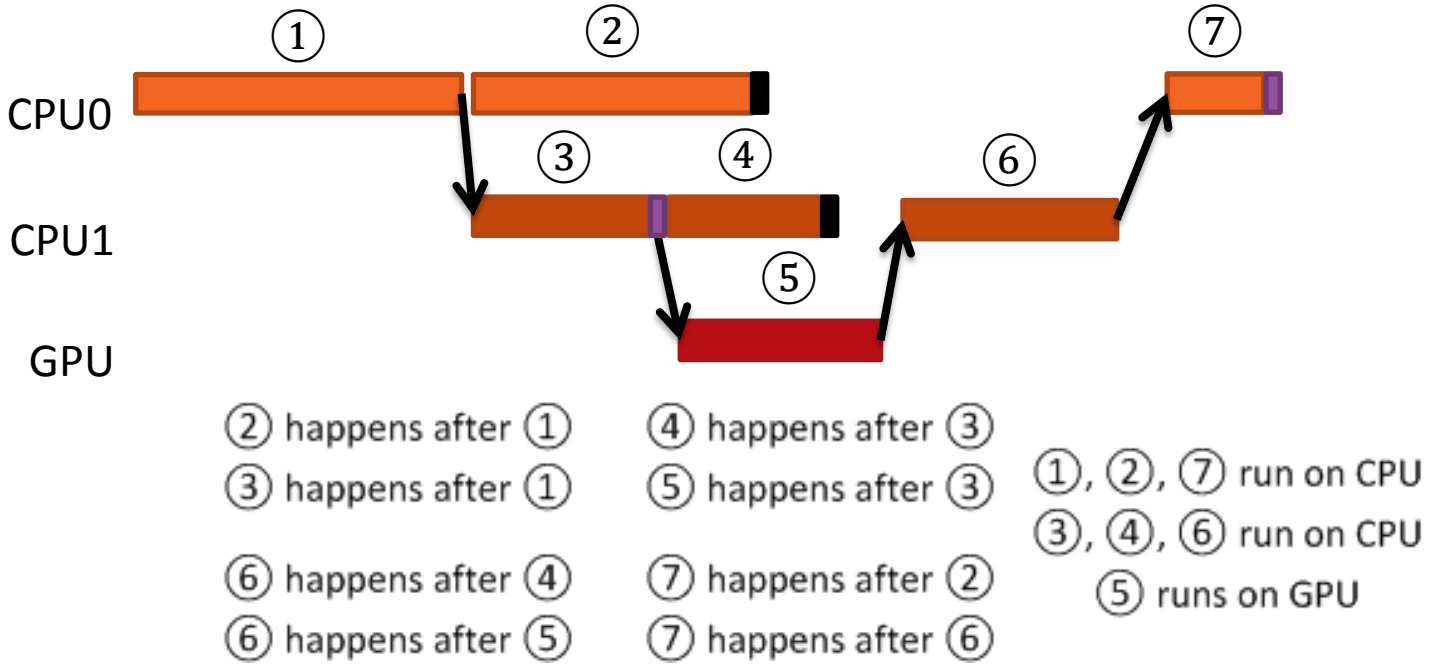


Example when everything except CPU1 gets faster:



MUST RECONSTRUCT CRITICAL PATHS

- Gather program-level relationship between individually scaled **segments**
- Use these happens-before relationships to build a legal execution order



- Gather ordering from library calls like pthread_create() and clWaitForEvents()
- Can also split segments based on program phases

CONCLUSION

AMD

Exascale node design space is huge

Trade off some accuracy for faster simulation

Use analytic models based on information from existing hardware

Research Related Questions

MODSIM RELATED QUESTIONS

AMD

- Major Contributions:
  - A fast, high-level performance, power, and thermal analysis infrastructure
  - Enables large design space exploration and HW/SW co-design with good feedback

- Limitations:
  - Trace-based simulation has known limitations w/r/t multiple paths of execution, wrong-path operations, etc.
  - It can be difficult and slow to model something if your hardware can't measure values that are correlated to it.

- Bigger Picture:
  - Node-level performance model for datacenter/cluster performance modeling
  - First pass model for APU power sharing algorithms.
  - Exascale application co-design
  - Complementary work to broad projects like SST

MODSIM RELATED QUESTIONS

AMD

- What is the one thing that would make it easier to leverage the results of other projects to further your own research
  - Theoretical bounds and analytic reasoning behind performance numbers. Even “good enough” guesses may help, vs. only giving the output of a simulator

- What are important thing to address in future work?
  - Better analytic scaling models. There are a lot in the literature, but many rely on simulation to propose new hardware that would gather correct statistics.
  - It would be great if open source performance monitoring software were better funded, had more people, etc.

DISCLAIMER & ATTRIBUTION

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

Backup

AN EXASCALE NODE EXAMPLE QUESTION

AMD

- Is 3D-Stacked Memory Beneficial for Application X?

Baseline Performance
Bandwidth Difference
Latency Difference
Thermal Model
Core Changes due to Heat

AN EXASCALE NODE EXAMPLE QUESTION

AMD

- Is 3D-Stacked Memory Beneficial for Application X?

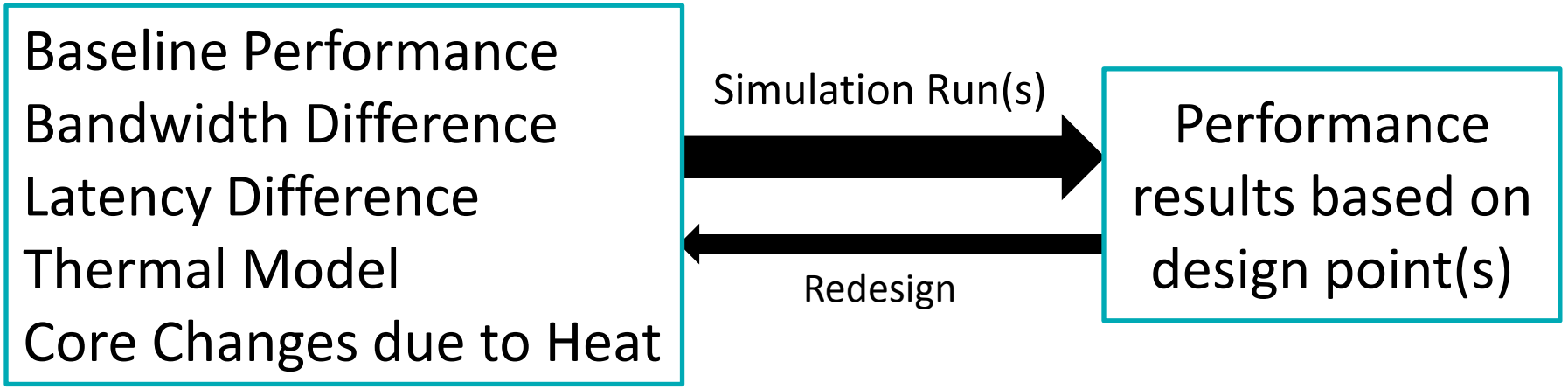
Baseline Performance
Bandwidth Difference
Latency Difference
Thermal Model
Core Changes due to Heat

Performance results based on design point(s)

AN EXASCALE NODE EXAMPLE QUESTION

AMD

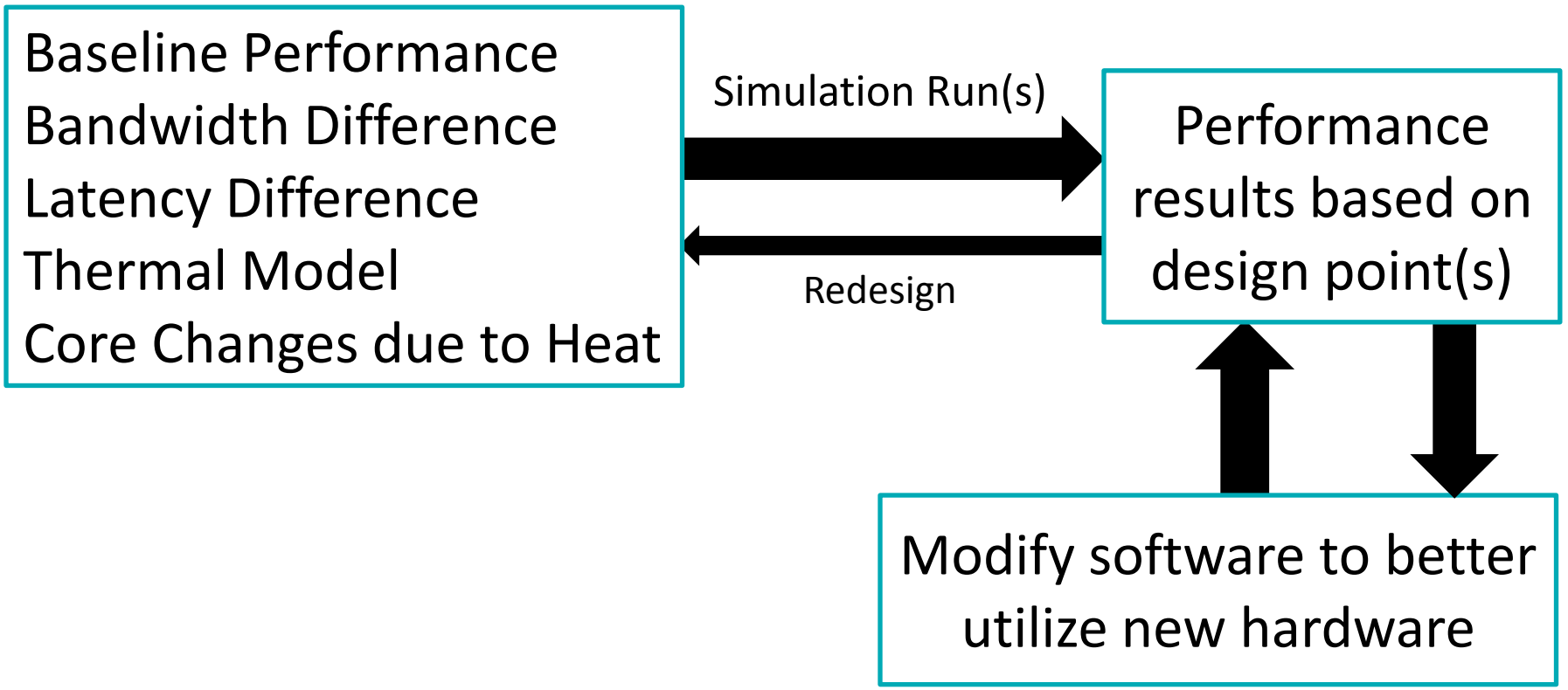
- Is 3D-Stacked Memory Beneficial for Application X?



AN EXASCALE NODE EXAMPLE QUESTION

AMD

- Is 3D-Stacked Memory Beneficial for Application X?



AN EXASCALE NODE EXAMPLE QUESTION

AMD

- Is 3D-Stacked Memory Beneficial for Application X?