HotGauge: A Methodology for Characterizing Advanced Hotspots in Modern and Next Generation Processors

Alexander Hankin[1*], **David Werner**[1*], Julien Sebot[2], Kaushik Vaidyanathan[3], Maziar Amiraski[1], Mark Hempstead[1]

[1]Tufts University, [2]Intel, [3]Google [*]co-authors

2021 IEEE International Symposium on Workload Characterization

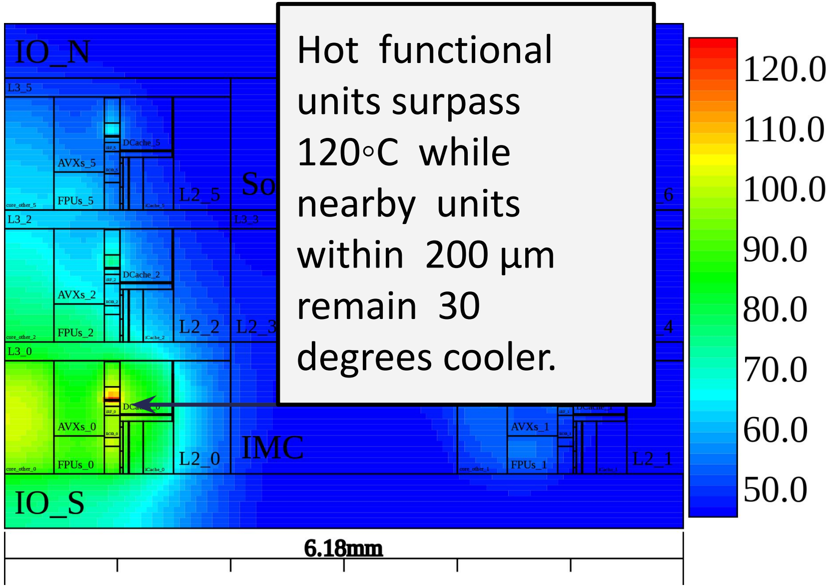
November 7, 2021

Tufts UNIVERSITY | SCHOOL OF ENGINEERING Electrical and Computer Engineering

Introduction

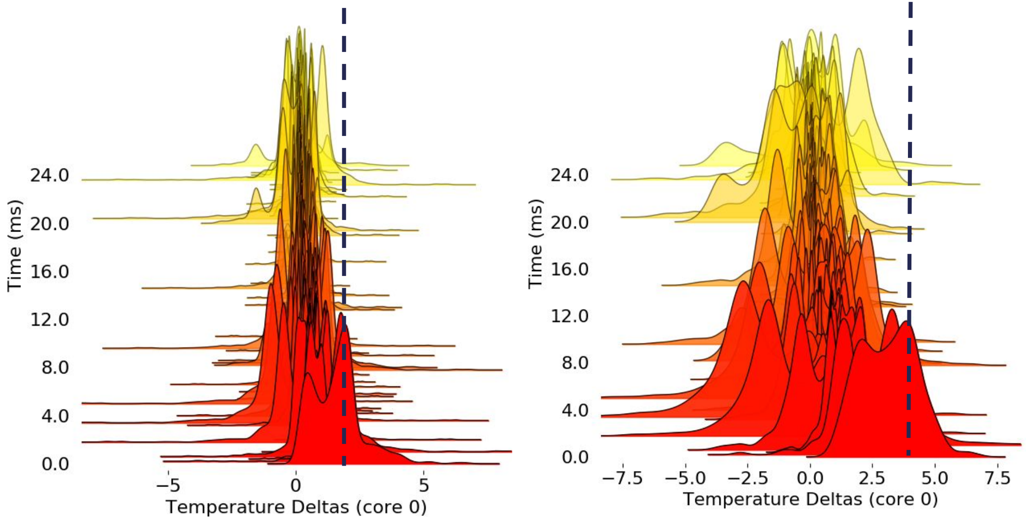
- System design trends have exacerbated thermal hotspots and have created “advanced” hotspots
  - Exponential power density
  - Cramming additional logic
- Advanced hotspots are a critical design concern and will continue to get worse without proper mitigation
- We have built the necessary tools for studying advanced hotspots and developing mitigation techniques

How bad are they?



Advanced hotspot in a 7nm client processor similar to Intel Skylake

How bad are they?



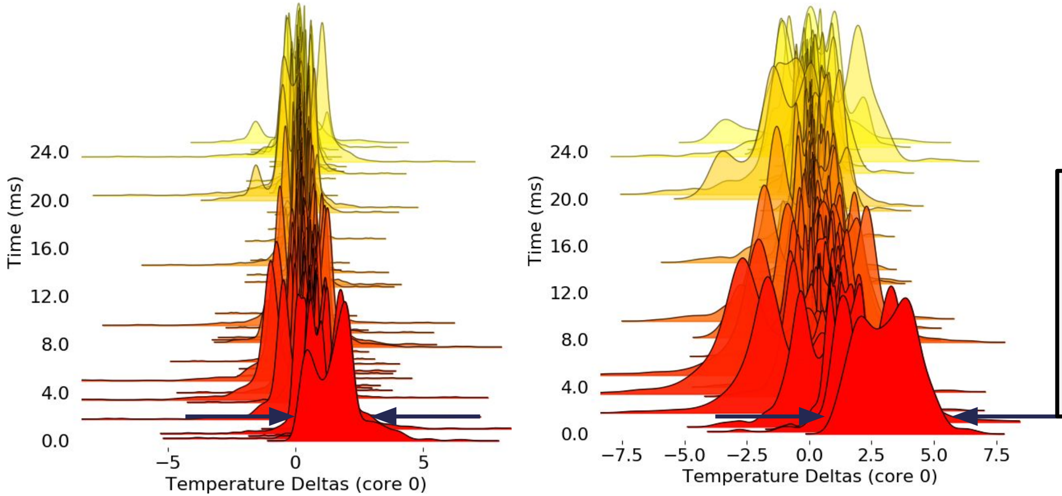
(a) 14nm

(b) 7nm

Distribution of temperature deltas over 200 µs intervals

Higher temperature deltas in 7nm.

How bad are they?



(a) 14nm (b) 7nm

Distribution of temperature deltas over 200 µs intervals

Wider distribution of temperature deltas within 200µs.

Background and Motivation

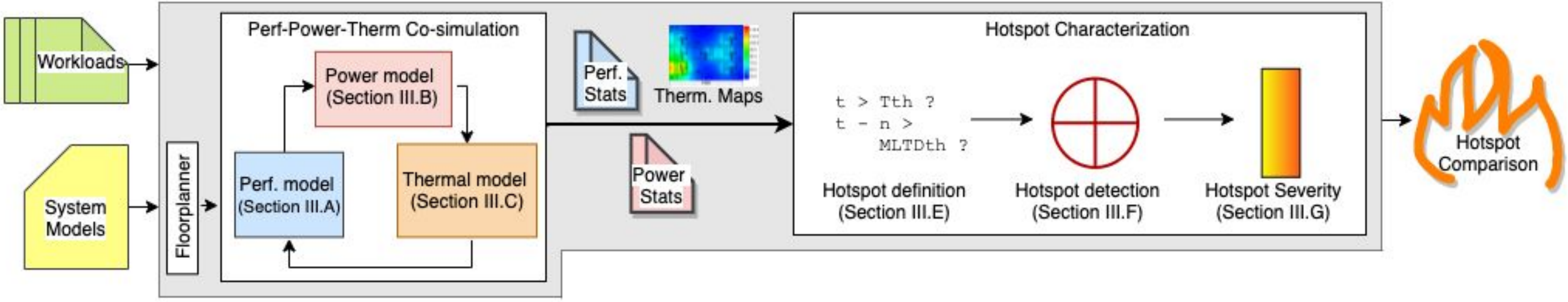
- Power density is localized, application-dependant, and rapidly increasing
  - standard thermal controllers won't work
- Lack of a hotspot definition
- Existing simulation methodologies are limited

Project Goal

TODO: sell the project

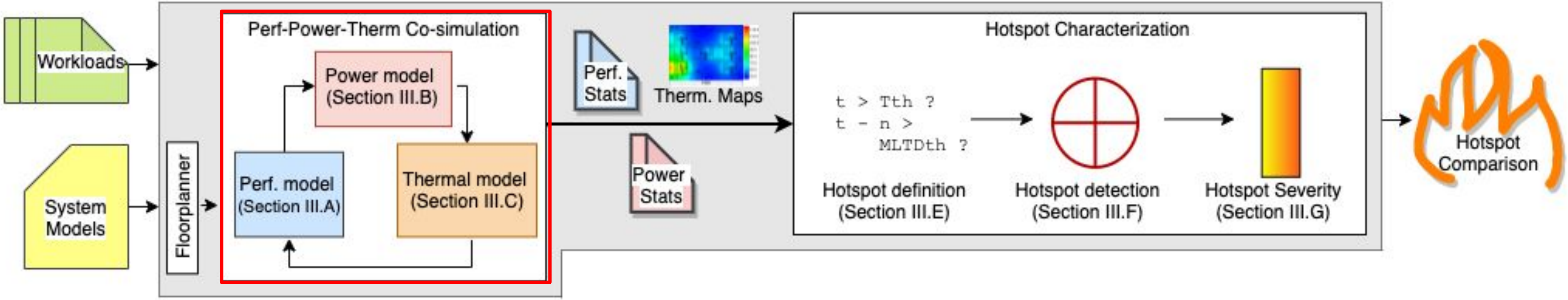
- Novel metrics and analysis techniques
- End-to-end simulation infrastructure
- Demonstrate using case studies

HotGauge



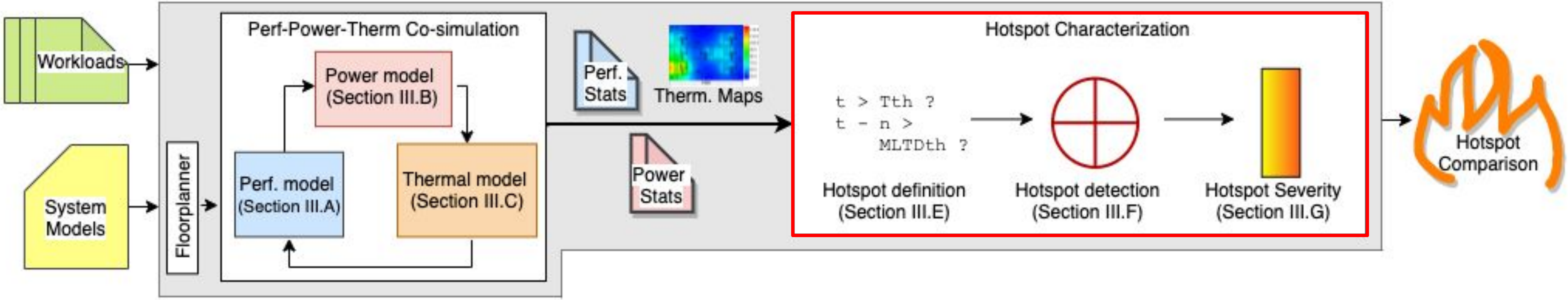
Workloads
System Models
Floorplanner
Perf-Power-Therm Co-simulation
Power model (Section III.B)
Perf. model (Section III.A)
Thermal model (Section III.C)
Perf. Stats
Therm. Maps
Power Stats
Hotspot Characterization
t > Tth ?
t - n > MLTDth ?
Hotspot definition (Section III.E)
Hotspot detection (Section III.F)
Hotspot Severity (Section III.G)
Hotspot Comparison

Perf:Power:Therm Simulation



- HotGauge takes as input the system models and workloads to be evaluated for hotspots
- First major task, Perf-power-therm co-simulation, integrates performance, power, and thermal models to perform rapid end-to-end thermal simulation

Hotspot Characterization

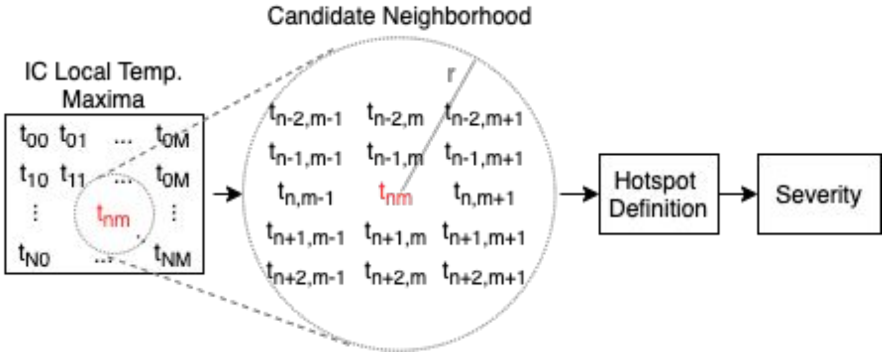


- Rigorous definition of a hotspot
- Automated method for detecting them
- Quantitative metric to compare hotspot severity between floorplans, or other model changes

Tufts UNIVERSITY | SCHOOL OF ENGINEERING Electrical and Computer Engineering

Hotspot Definition and Detection

- <u>Goal:</u> capture all phenomena which will cause either performance loss or reliability problems
  - absolute temperature
  - maximum localized temperature differential (MLTD)



Hotspot Severity Metric

- Hotspot Severity accounts for *both* MLTD and temperature
- Comprised of multiple parameterized sigmoid functions, each dealing with a different concern

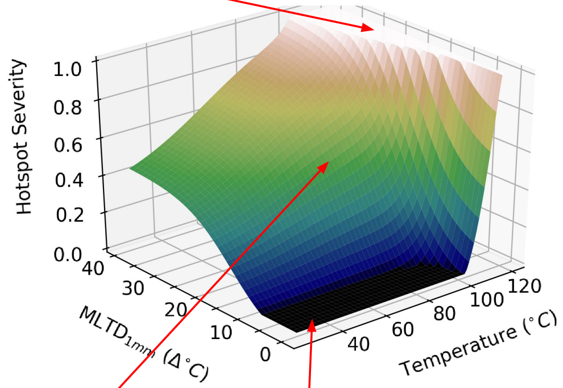
$$\sigma(x_o, y_o, s, a) = \frac{a}{1 + e^{-s(x - x_o)}} + y_o$$

$$sev(x,y) \;=\; \sigma_{df}(T_{x,y}) + \sigma_M(MLTD_{x,y}) * \sigma_T(T_{x,y})$$

where:

$$\begin{aligned}
\sigma_{df} &= \sigma(115, 0, 0.2, 2) \\
\sigma_M &= \sigma(15, -0.25, 0.2, 1.25) \\
\sigma_T &= \sigma(60, 0.35, 0.05, 0.65)
\end{aligned}$$

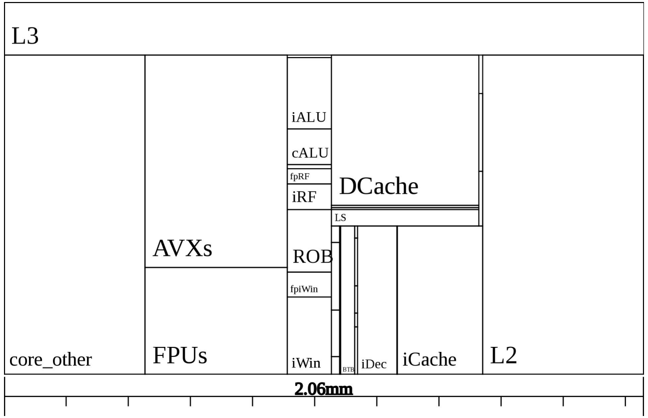
1 → crash or permanent damage imminent



0.5 → immediate mitigation required

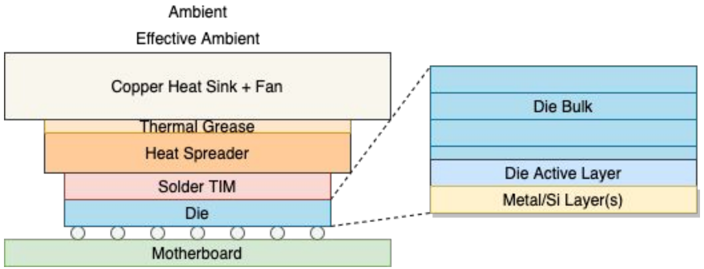
0 → no concern

Case Study - Skylake Proxy



CPU Microarchitecture Parameters	
Process node [nm]	14, 10, 7
Cores	7
Core area [mm^2]	5, 2.5, 1.25
Frequency	5 GHz
SMT	2
ROB entries	224
LQ entries	72
SQ entries	56
Scheduler entries	97
L1I $	Private, 32 KiB
L1D $	Private, 32 KiB
L2 $	Private, 512 KiB
L3 $	Shared ring, 16 MiB

Layer	Thermal Conductivity [W/μm K]	Volumetric Heat Capacity [J/μm^3 K]	height [μm]
Thermal grease	0.04e-4	3.376e-12	30
Copper (heat spreader)	3.9e-4	3.376e-12	3e3
Solder TIM	0.25e-4	1.628e-12	200
Silicon (IC wafer)	1.20e-4	1.651e-12	380



Evaluation and Case Study

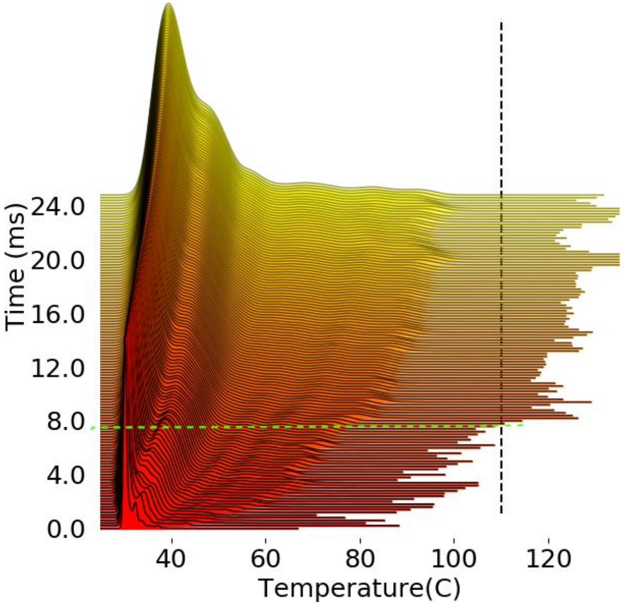
Using *HotGauge*, we aim to...

- Quantify how bad the problem is
- How much worse will the problem be in next generation processors
- Evaluating mitigations

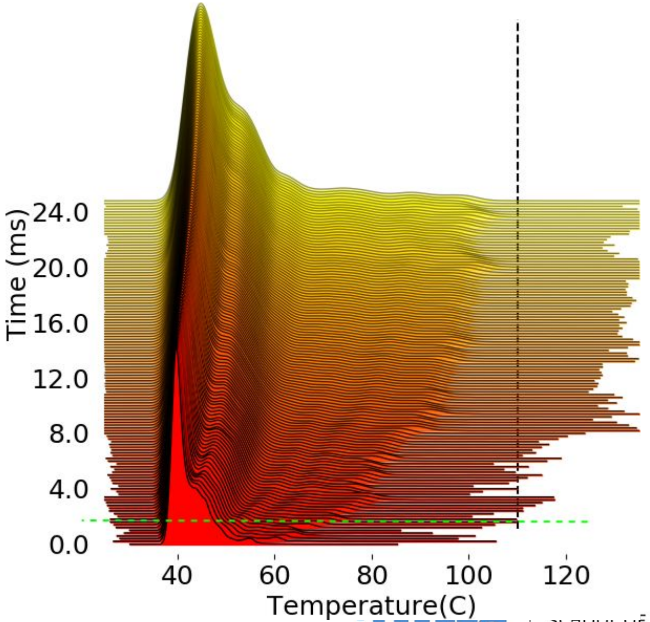
Tufts UNIVERSITY | SCHOOL OF ENGINEERING Electrical and Computer Engineering

Results - Temperature Distribution vs Time

Ambient



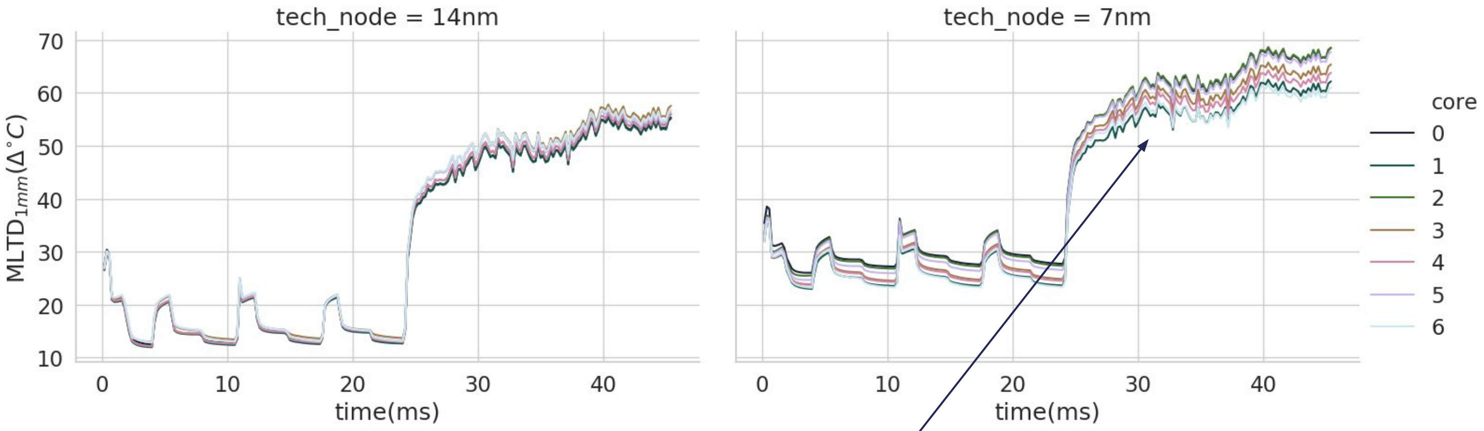
Idle



Initial conditions dramatically affect **how quickly hotspots can arise** and **distribution of IC temperatures**

Tufts University | School of Engineering
Electrical and Computer Engineering

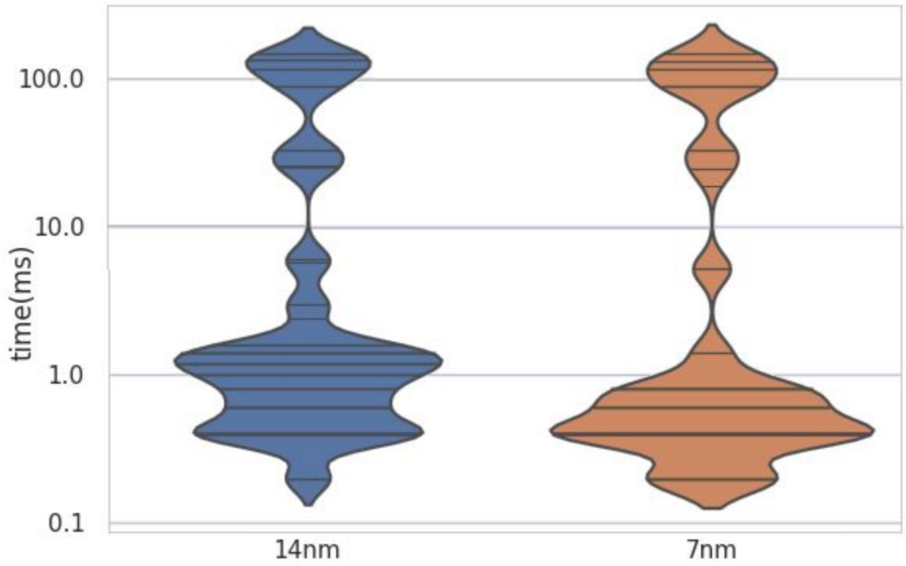
MLTD vs. Tech Node



Higher MLTD in newer 7nm node

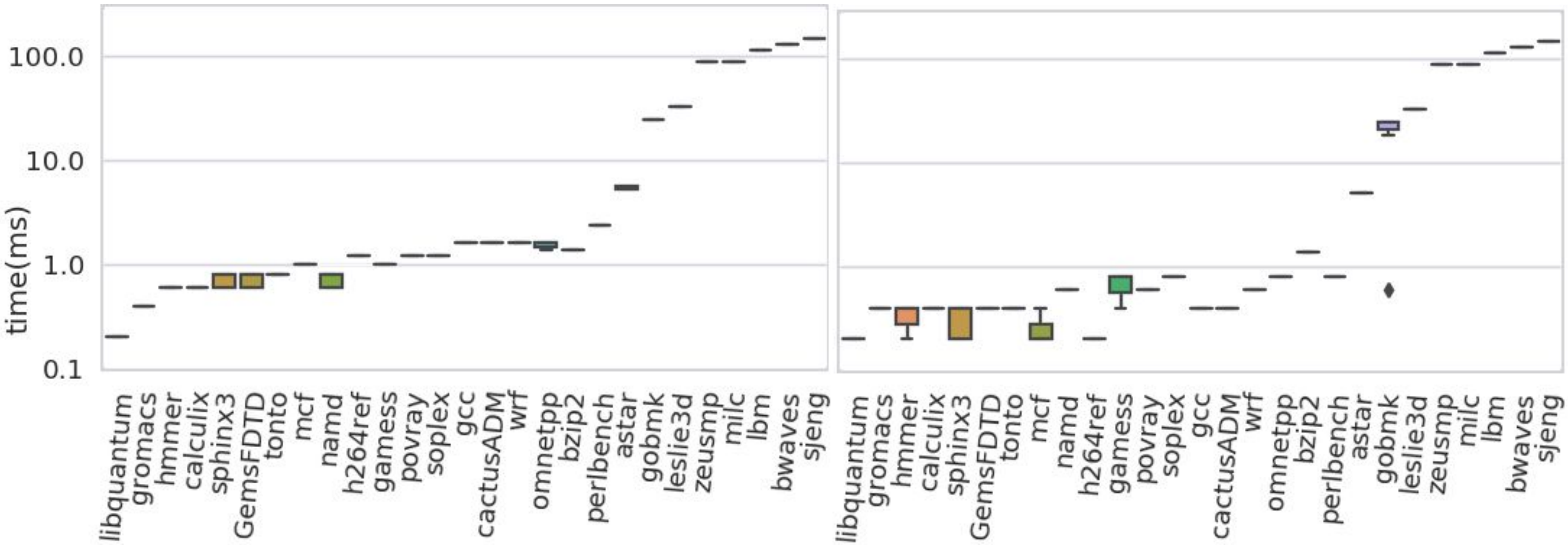
MLTD varies more from core to core

Results -- TUH vs tech node



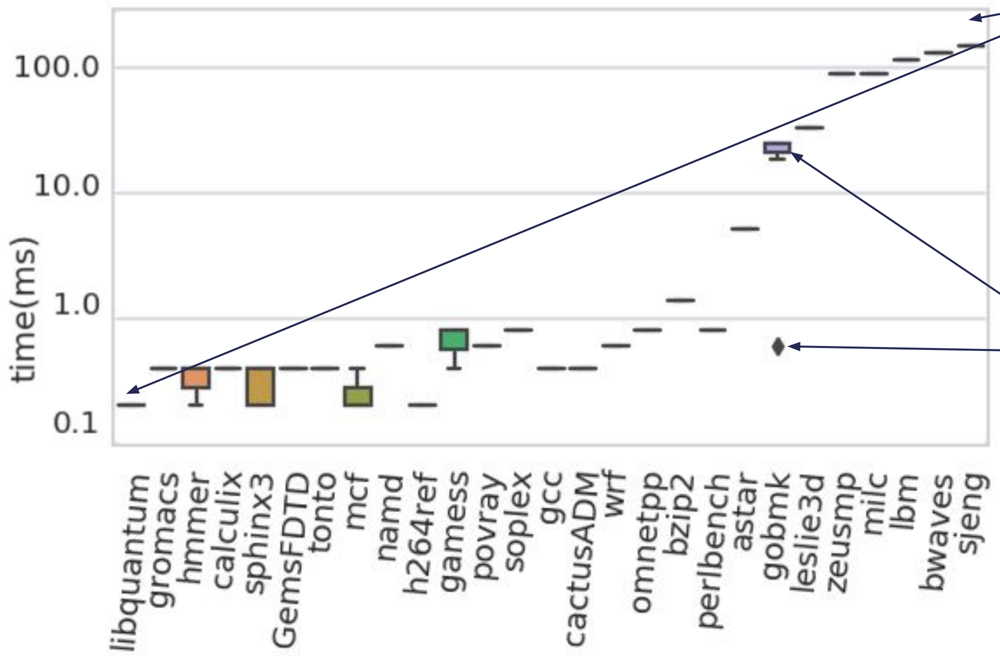
Tufts UNIVERSITY | SCHOOL OF ENGINEERING Electrical and Computer Engineering

Results -- TUH vs warmup



SCHOOL OF ENGINEERING
Electrical and Computer Engineering

Time Until Hotspot

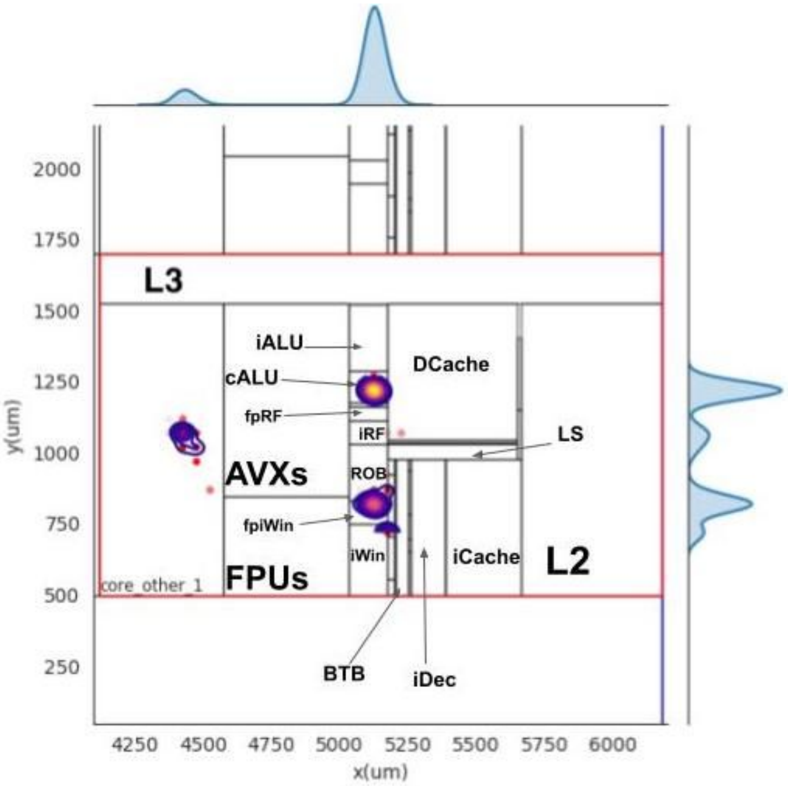


Hotspot behaviour across benchmarks varies considerably

Some benchmarks are more affected by core-placement

Tufts UNIVERSITY | SCHOOL OF ENGINEERING Electrical and Computer Engineering

Hotspot Locations



Some areas are prone to hotspots

Some areas have infrequent hotspots

Tufts UNIVERSITY | SCHOOL OF ENGINEERING Electrical and Computer Engineering

Case Study
Single Unit Mitigation

Goal : decrease hotspots caused by *problematic* units

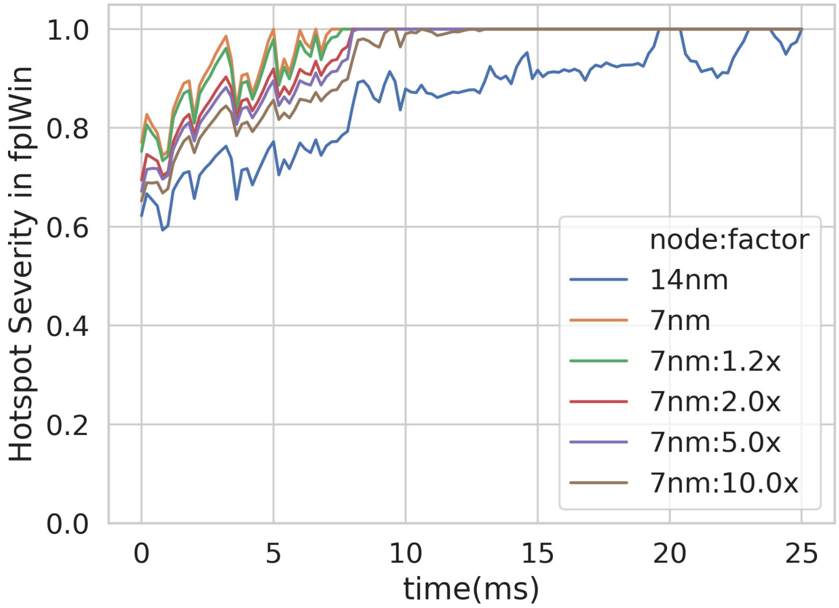
Baseline : make 7nm unit no worse than 14nm unit
(using hotspot-severity)

Method : scale area of problematic units
(thereby decreasing power-density)

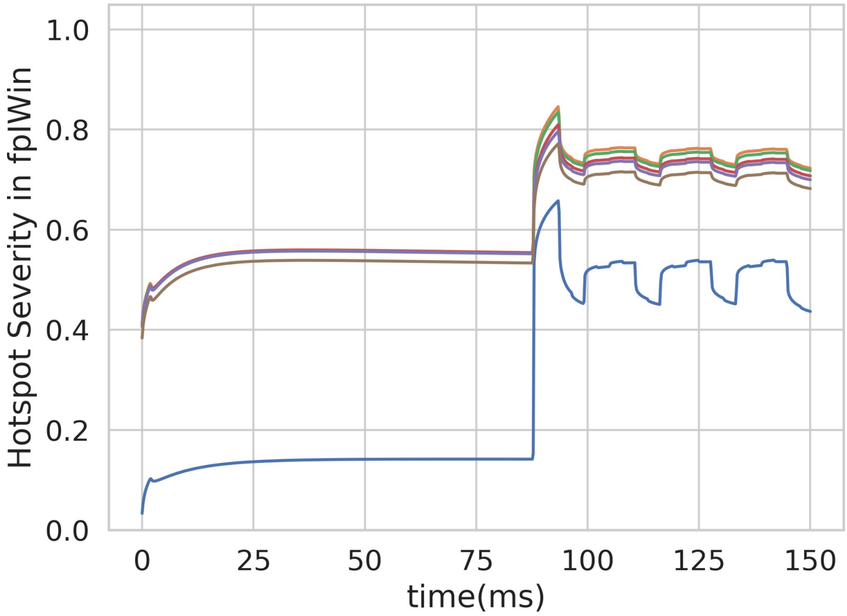
Tufts UNIVERSITY | SCHOOL OF ENGINEERING Electrical and Computer Engineering

Unit Scaling

Floating Point Instruction Window



gcc



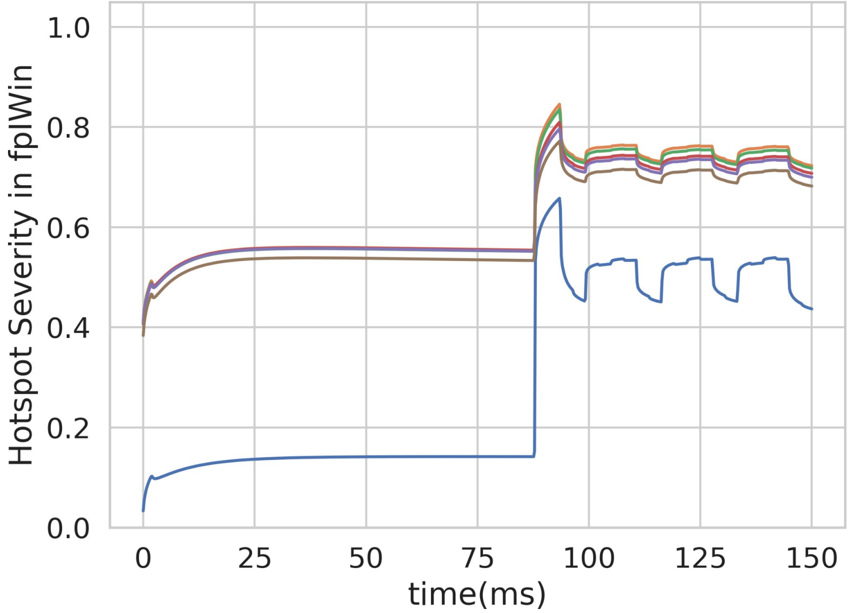
milc

Effectiveness of unit scaling varies from benchmark to benchmark

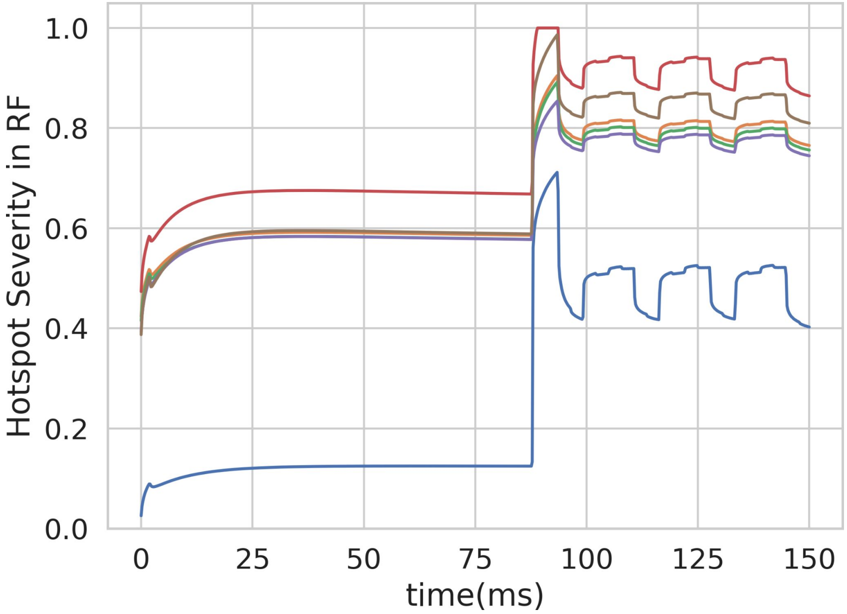
Tufts UNIVERSITY | SCHOOL OF ENGINEERING Electrical and Computer Engineering

Unit Scaling

Different Units for *milc*



FP Inst. Window

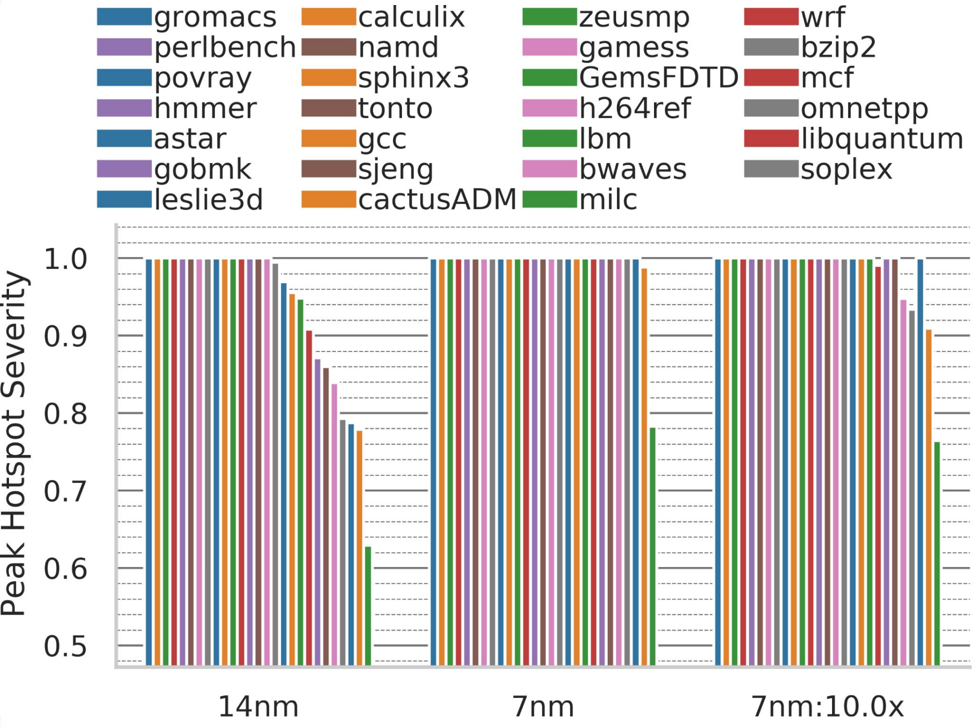


Register Files

Effectiveness of unit scaling varies from benchmark to benchmark

Tufts UNIVERSITY | SCHOOL OF ENGINEERING Electrical and Computer Engineering

Scaling Register Access Tables



- Hotspot severity of 1.0:
  - 14nm: few workloads
  - 7nm: most workloads
- Varied impact of scaling RATs across workloads

Other Contributions

- Model validation
  - Thermal stack
  - Power Model
- Modified Sniper/MCPAT for new nodes
- Case Study: Scaling the entire IC

Conclusion

- In this work we introduce HotGauge, a holistic methodology for characterizing hotspots in modern and next generation processors

- HotGauge details new methods and metrics for characterizing and comparing hotspot severity across any next generation processors.

- This will allow the architecture community to develop architecture level mitigations to work alongside traditional thermal regulation techniques to solve the advanced thermal hotspots which are occurring in modern and next generation processors.

HotGauge FrameWork

Publicly available on GitHub!

TuftsCompArchLab / **HotGauge** Public

https://github.com/TuftsCompArchLab/HotGauge

Includes Docker container to run thermal-simulations and perform analysis

docker

Tufts UNIVERSITY | SCHOOL OF ENGINEERING Electrical and Computer Engineering

HotGauge: A Methodology for Characterizing Advanced Hotspots in Modern and Next Generation Processors

Alexander Hankin[1*], David Werner[1*], Julien Sebot[2], Kaushik Vaidyanathan[3], Maziar Amiraski[1], Mark Hempstead[1]

[1]Tufts University, [2]Intel, [3]Google [*]co-authors

https://github.com/TuftsCompArchLab/HotGauge

Tufts UNIVERSITY | SCHOOL OF ENGINEERING Electrical and Computer Engineering

Validation

	C_{dyn} [nF]					
	14nm Si	model	error	10nm Si	model	error
bzip2	1.33	1.36	+2%	1.32	1.17	-11%
gcc	1.51	1.30	-14%	1.80	1.13	-37%
omnetpp	1.16	1.33	+15%	0.99	1.16	+17%
povray	1.87	1.62	-13%	1.87	1.36	-27%
hmmer	1.52	1.65	+9%	1.49	1.38	-7%
abs. avg.	-	-	**11%**	-	-	**20%**

	Process Node [nm]		
	14	10	7
Ψ [$^\circ C$/W]	0.96	1.13	1.40
TDP [W]	63	53	43