Safe & Effective Overclocking

Sk Hasibul Alam

ECE 651 – CAD of VLSI Systems

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Outline

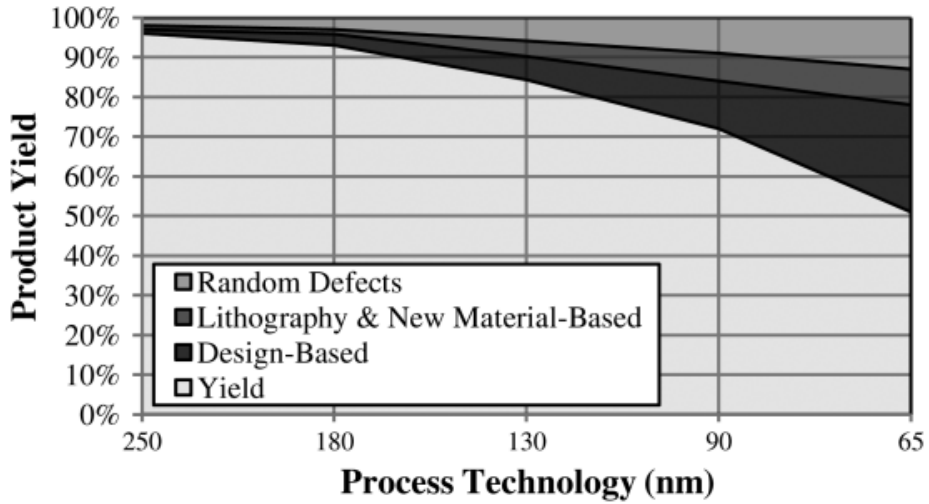
- Types of Overclocking
  - Effects on CPU
- Memory
- GPU
- Cooling
  - Datacenters
  - Smartphones
- Conclusion

What is overclocking?

Deliberate practice of exceeding a component's manufacturer-specified clock rate

Done by modifying:

- Clock multiplier
- Bus clock rate



Types of Overclocking

VF-Overclocking

- Increases f_{clk} and $\mathrm{V_{DD}}$ simultaneously
- Leads to very high $\mathrm{P}_{dynamic}$
- Examples:
  - Intel Turbo Boost, AMD Turbo Core

F-Overclocking

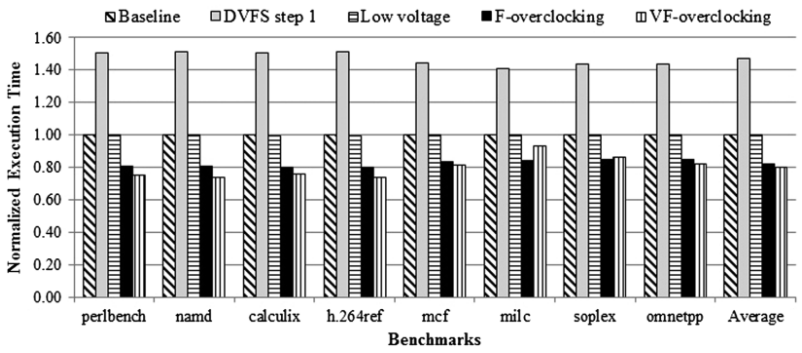
- Increases f_{clk} only
- Leads to high $\mathrm{P}_{dynamic}$
- Not widely adopted

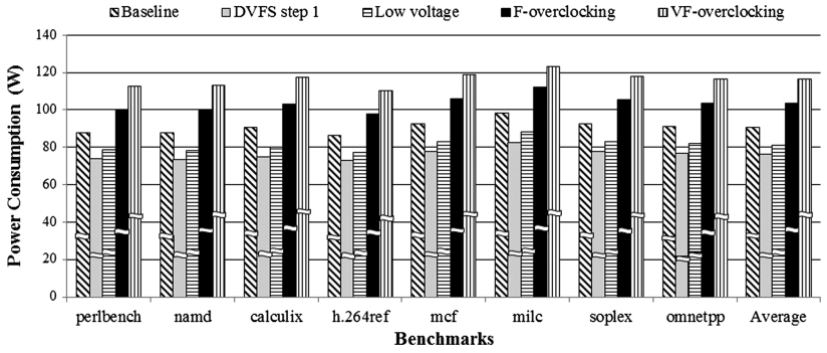
$$P_{total} = \alpha \mathrm{C} {\mathrm{V_{DD}}}^2 f_{clk} + P_{static}$$

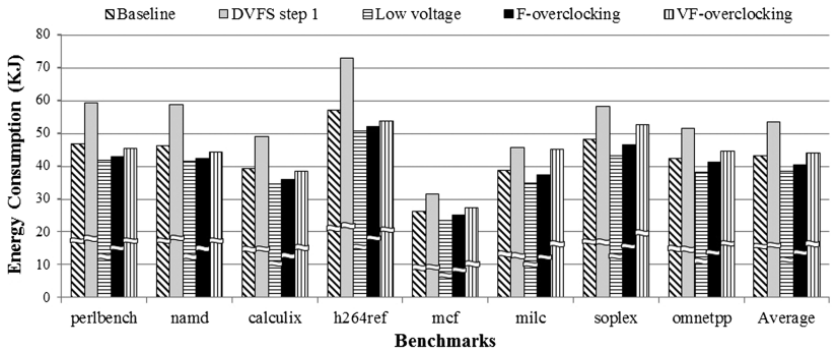
Evaluation of F-Overclocking

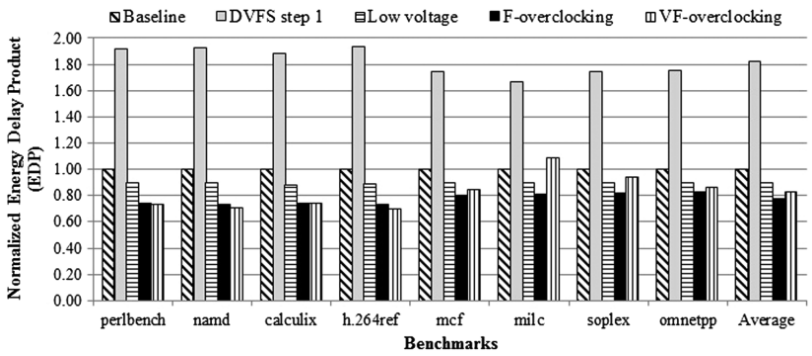
Scheme	Clock Frequency (MHz)	Supply Voltage (V)
Baseline	3000	1.250
DVFS Step 1	2000	1.100
Low Voltage	3000	1.075
VF-overclocking	4050	1.350
F-overclocking	3735	1.250

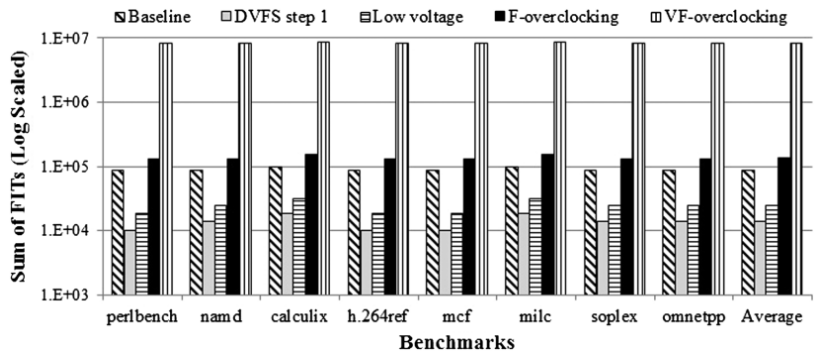
Platform: Intel Core 2 Duo E8400 (45 nm)

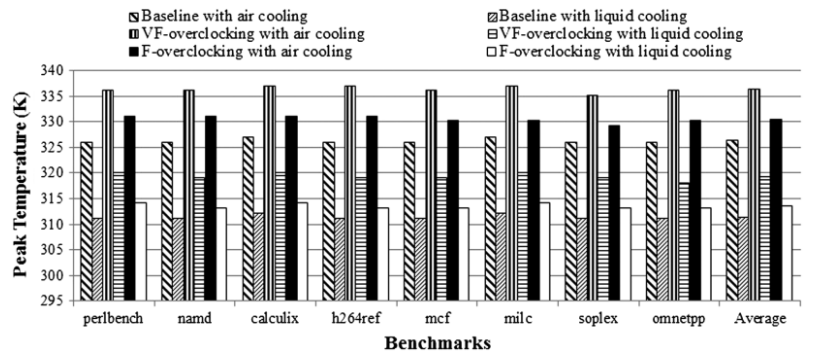






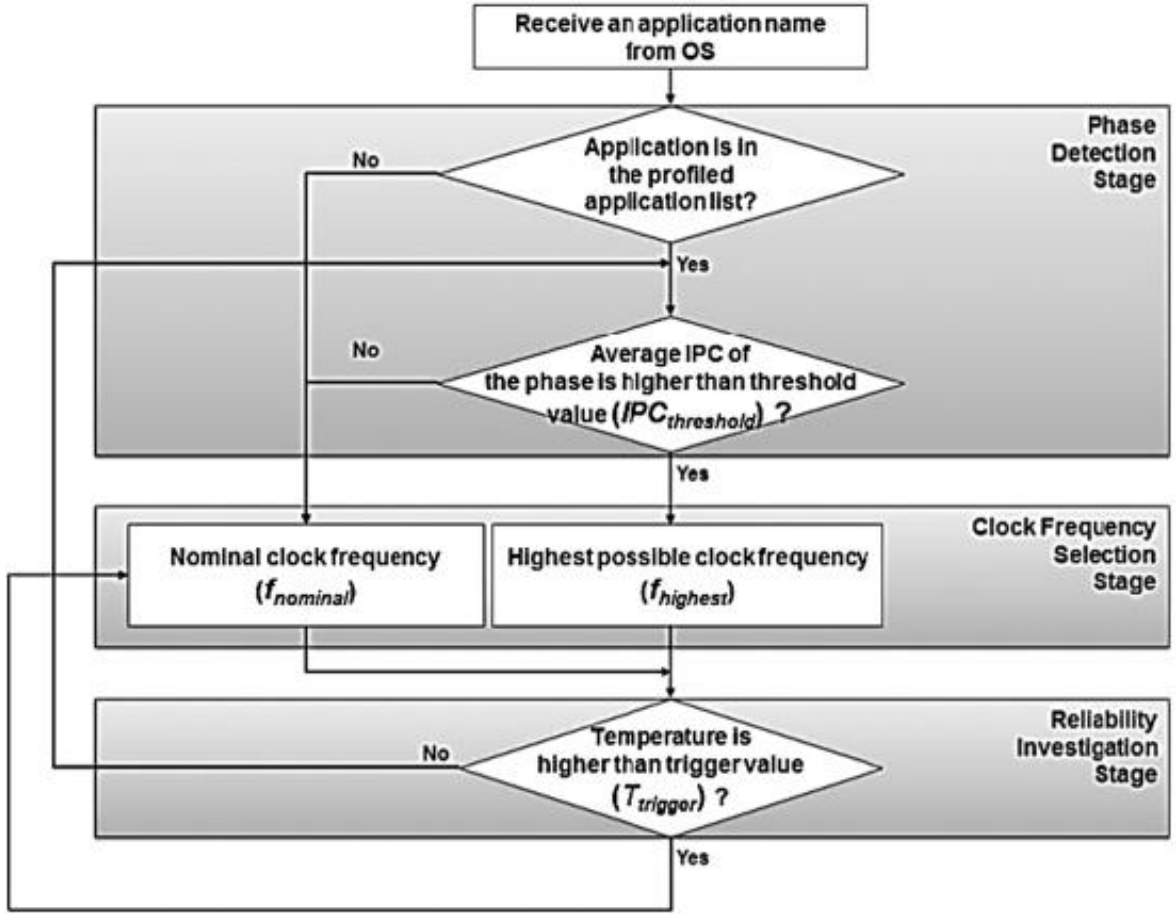






IEEE Transactions on Computers, 2014, DOI: 10.1109/TC.2012.286

Adaptive Overclocking Controller

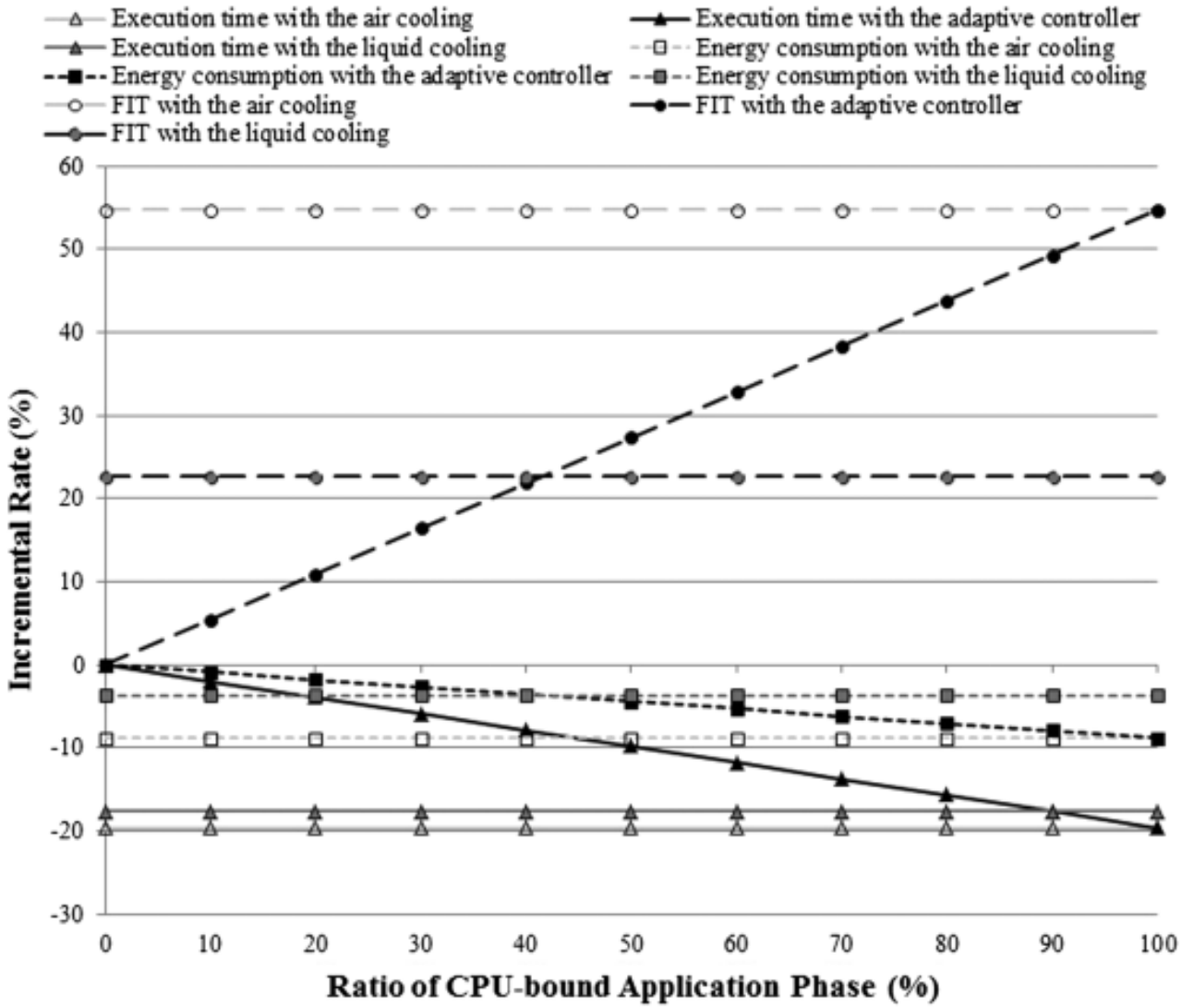


- For performance, set $IPC_{threshold}$ low.
- For reliability, set $IPC_{threshold}$ high.

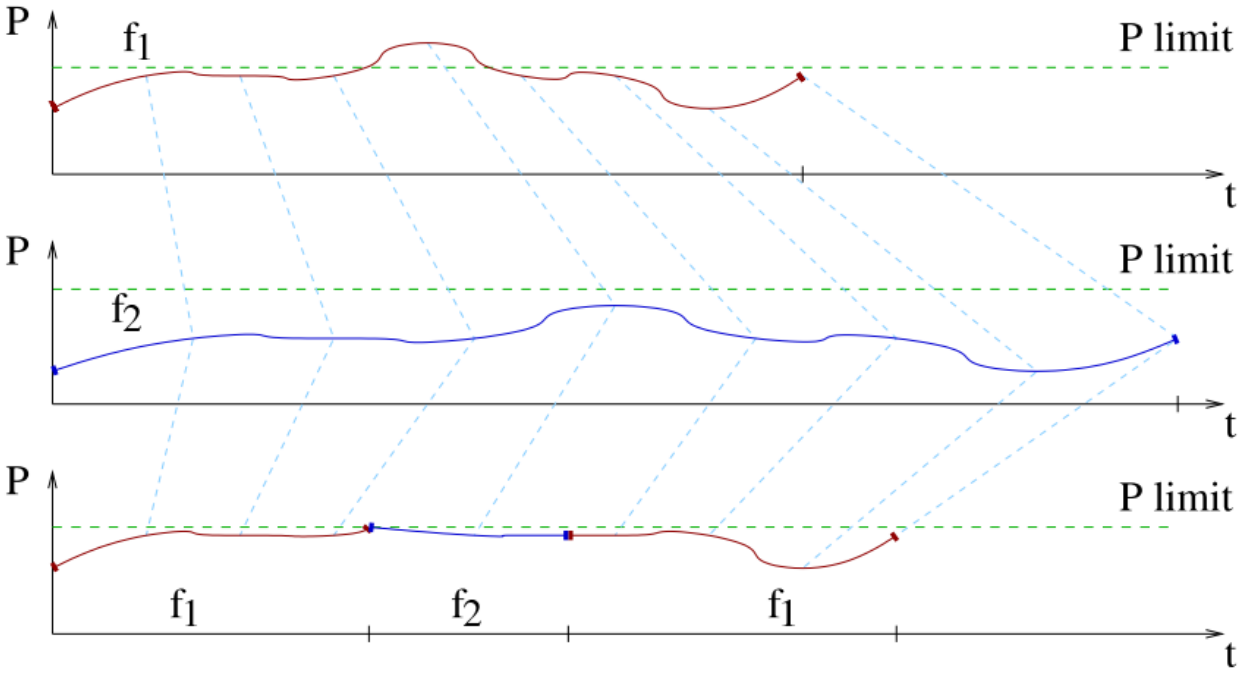
continued…

3 cases:

- F-Overclocking *with* air cooling
- F-Overclocking *with* liquid cooling
- F-Overclocking *with* adaptive controller & air cooling



Dynamic Processor Overclocking

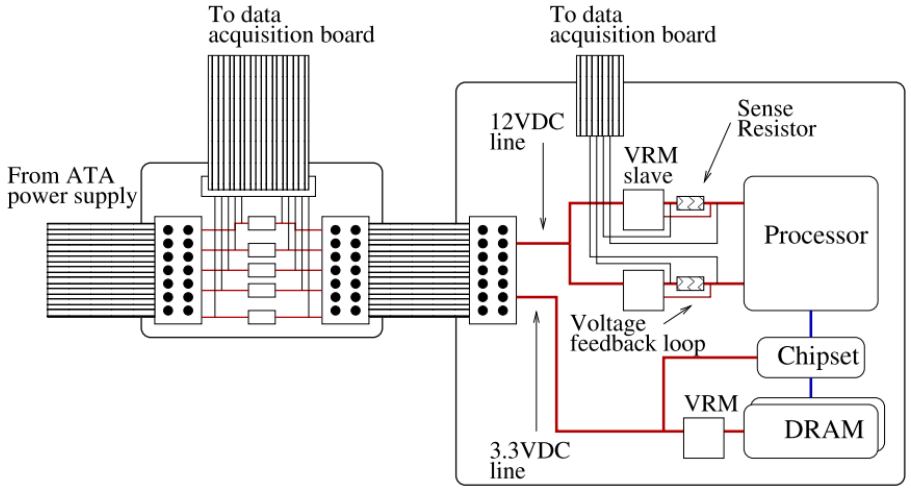


$f_1 > f_2$

A service is invoked every 10 ms.

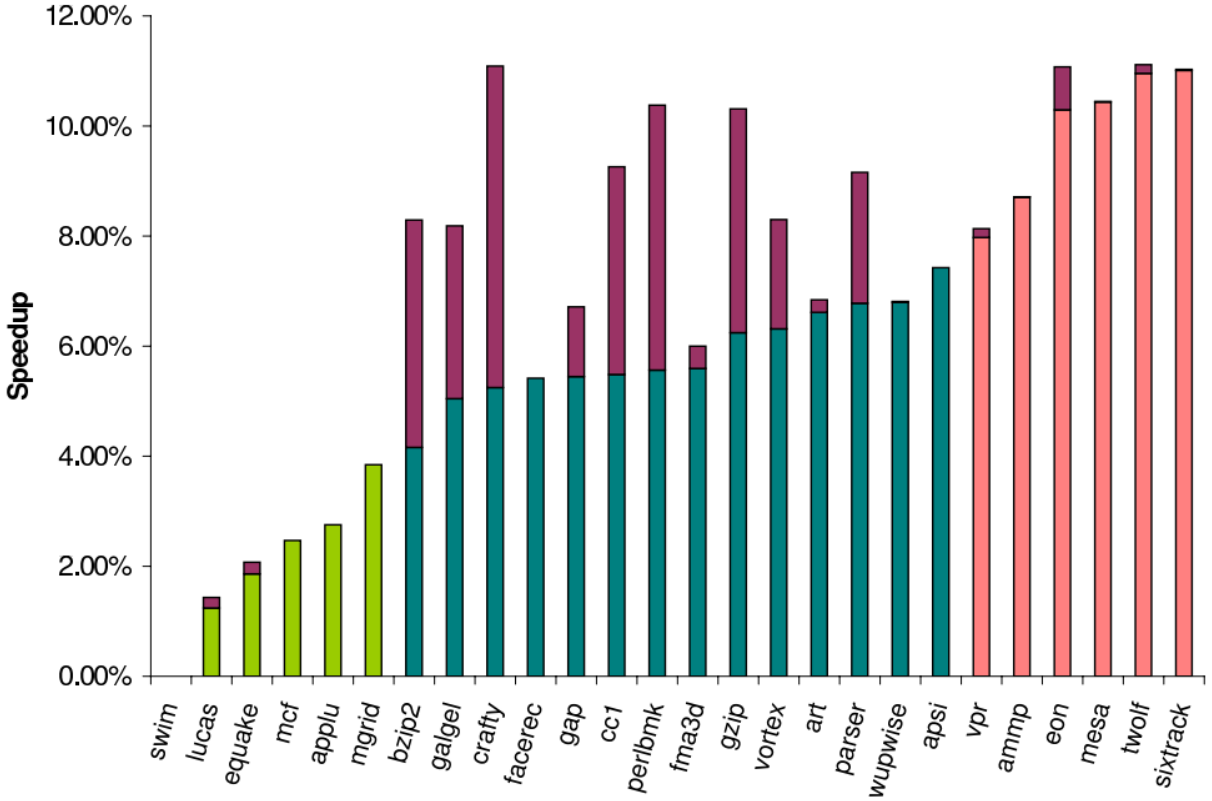
Monitor:

- Performance counters (Instr. Decode)
- Power

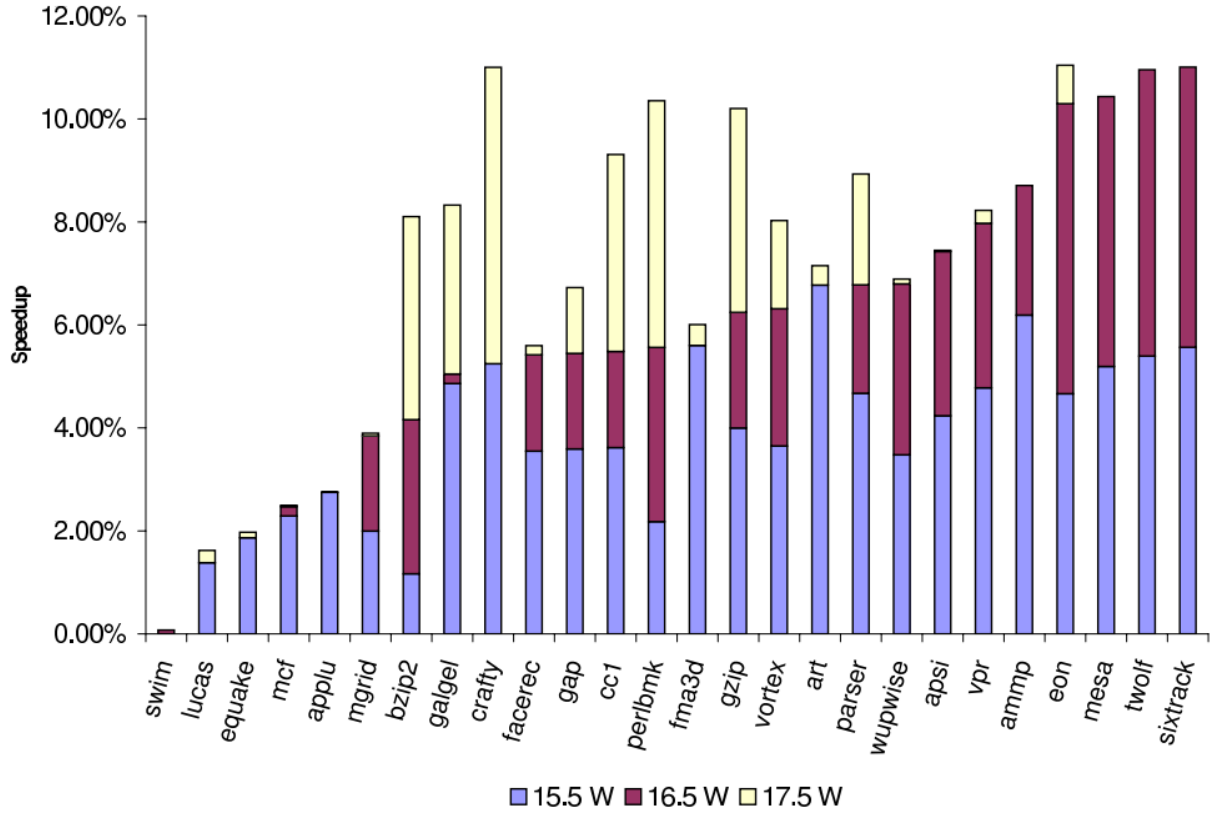


IBM Research Report RC23747, 2005, Austin, TX

continued…



Lower bar: DPO with 16.5 W power limit
Upper bar: 2.0 GHz with no power limit

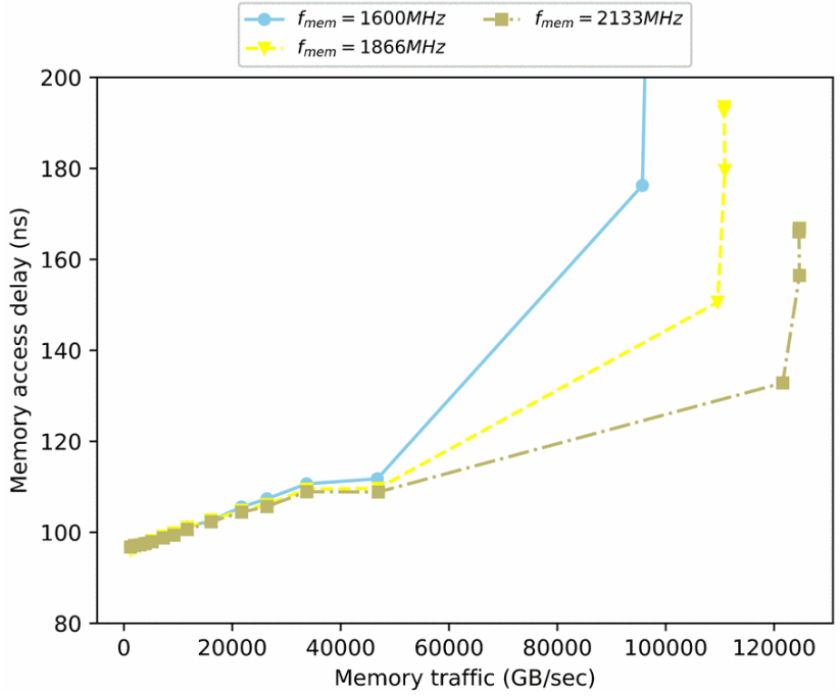


DPO with different power limits

Platform: Intel Pentium M (90 nm); $f_{nominal} = 1.8$ GHz

Memory Underclocking

- Memory power consumption: background + operation + R/W + I/O
- If memory traffic is high → CPU overclocking will degrade EDP.
- If EDP improves from CPU overclocking → memory traffic is low.
  - ➢ Memory underclocking will further improve EDP.



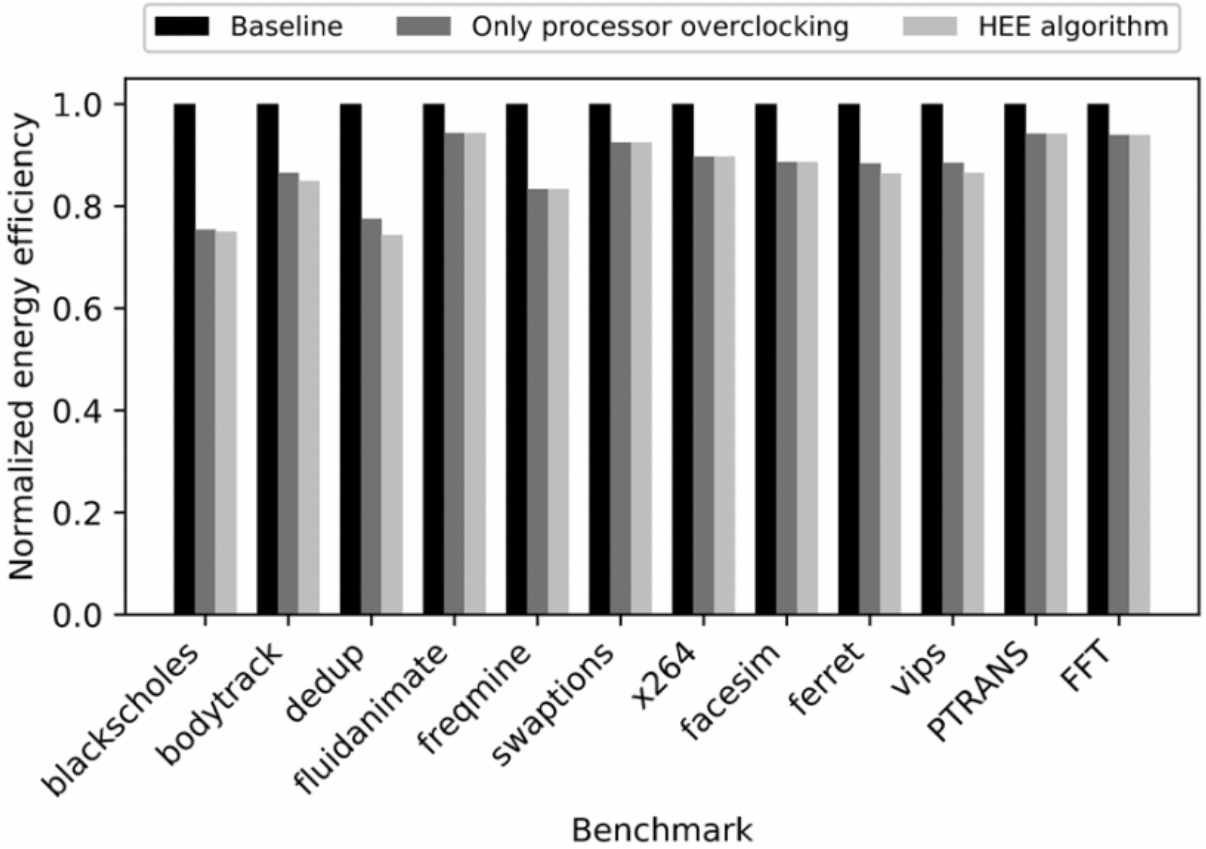
IEEE 4th Int. Conf. on Data Science and Systems, UK, 2018, DOI: 10.1109/HPCC/SmartCity/DSS.2018.00159

Holistic Energy-Efficient Algorithm

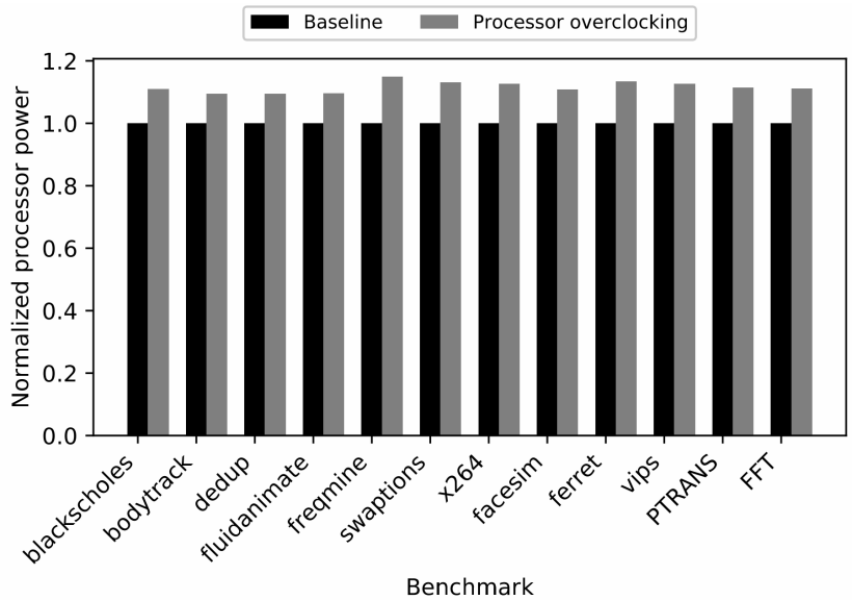
Input: A given program, memory traffic ratio threshold α, and processor time ratio threshold β
Output: is_{turbo}, $is_{scalemem}$
Obtain parameter $processor_{time\_ratio}$;
Obtain parameter memory traffic $traffic$;
$is_{turbo} = 0$;
$is_{scalemem} = 0$;
if $(processor_{time\_ratio} > \beta)$ **then**
    /* The program has a high processor time proportion */
    $is_{turbo} = 1$;
    if $(traffic < \alpha)$ **then**
        $is_{scalemem} = 1$;
    end
end

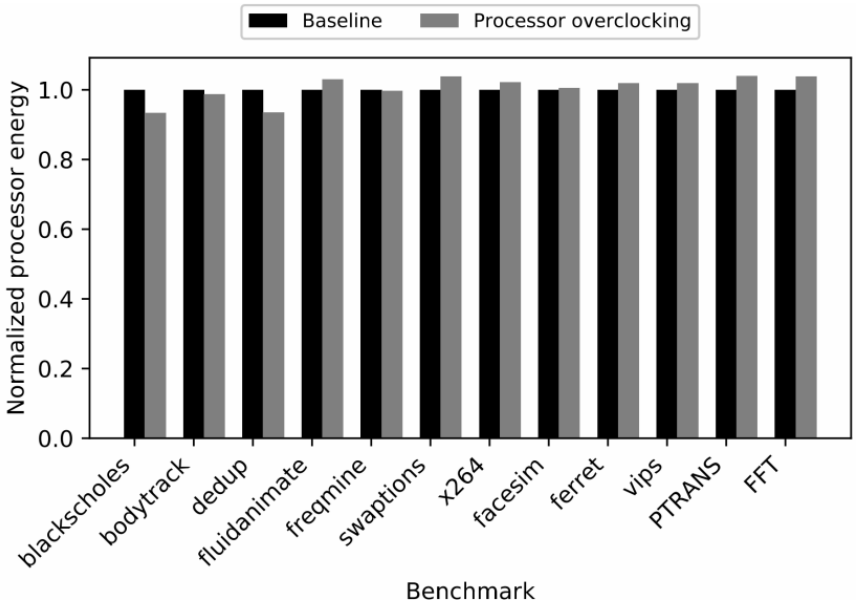
continued…

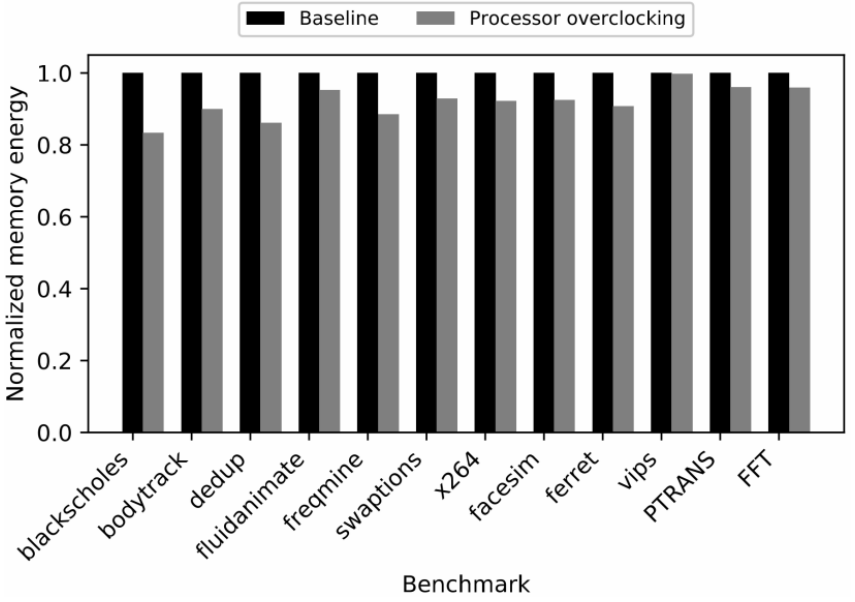
Benchmark	memory traffic	processor time ratio	is_{turbo}	$is_{scalemem}$
blackscholes	14	0.993	1	1
bodytrack	5	0.991	1	1
dedup	11	0.963	1	1
fluidanimate	38	0.925	1	0
freqmine	44	0.991	1	0
swaptions	17	0.971	1	0
x264	17	0.907	1	0
canneal	20	0.47	0	0
facesim	17	0.948	1	0
ferret	14	0.932	1	1
streamcluster	32	0.43	0	0
vips	5	0.992	1	1
RandomAccess	116	0.129	0	0
STREAM	104	0.637	0	0
FFT	113	0.851	1	0
PRANTS	107	0.877	1	0

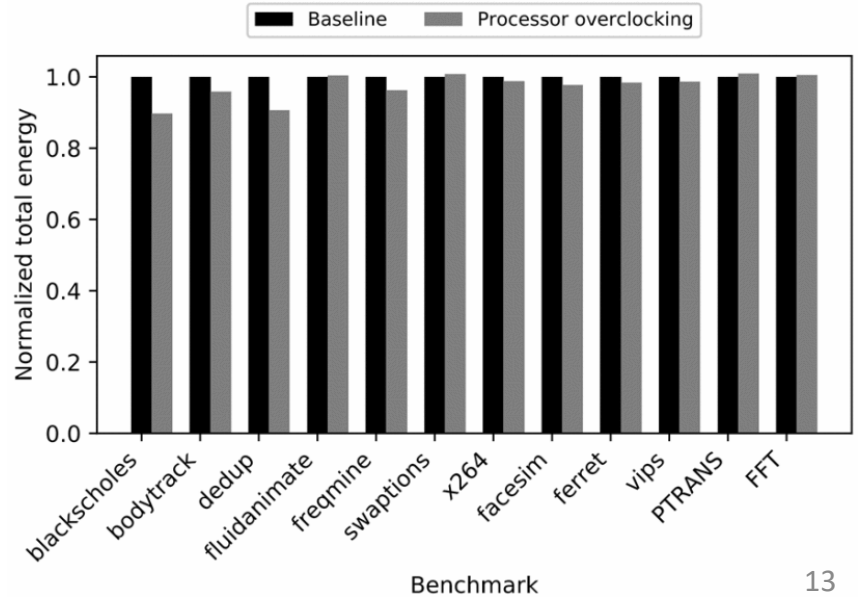


continued…


Baseline
Processor overclocking
Normalized processor power
1.2
1.0
0.8
0.6
0.4
0.2
0.0
blackscholes
bodytrack
dedup
fluidanimate
freqmine
swaptions
x264
facesim
ferret
vips
PTRANS
FFT
Benchmark


Baseline
Processor overclocking
Normalized processor energy
1.0
0.8
0.6
0.4
0.2
0.0
blackscholes
bodytrack
dedup
fluidanimate
freqmine
swaptions
x264
facesim
ferret
vips
PTRANS
FFT
Benchmark


Baseline
Processor overclocking
Normalized memory energy
1.0
0.8
0.6
0.4
0.2
0.0
blackscholes
bodytrack
dedup
fluidanimate
freqmine
swaptions
x264
facesim
ferret
vips
PTRANS
FFT
Benchmark


Baseline
Processor overclocking
Normalized total energy
1.0
0.8
0.6
0.4
0.2
0.0
blackscholes
bodytrack
dedup
fluidanimate
freqmine
swaptions
x264
facesim
ferret
vips
PTRANS
FFT
Benchmark

CPU-GPU Heterogenous Platforms

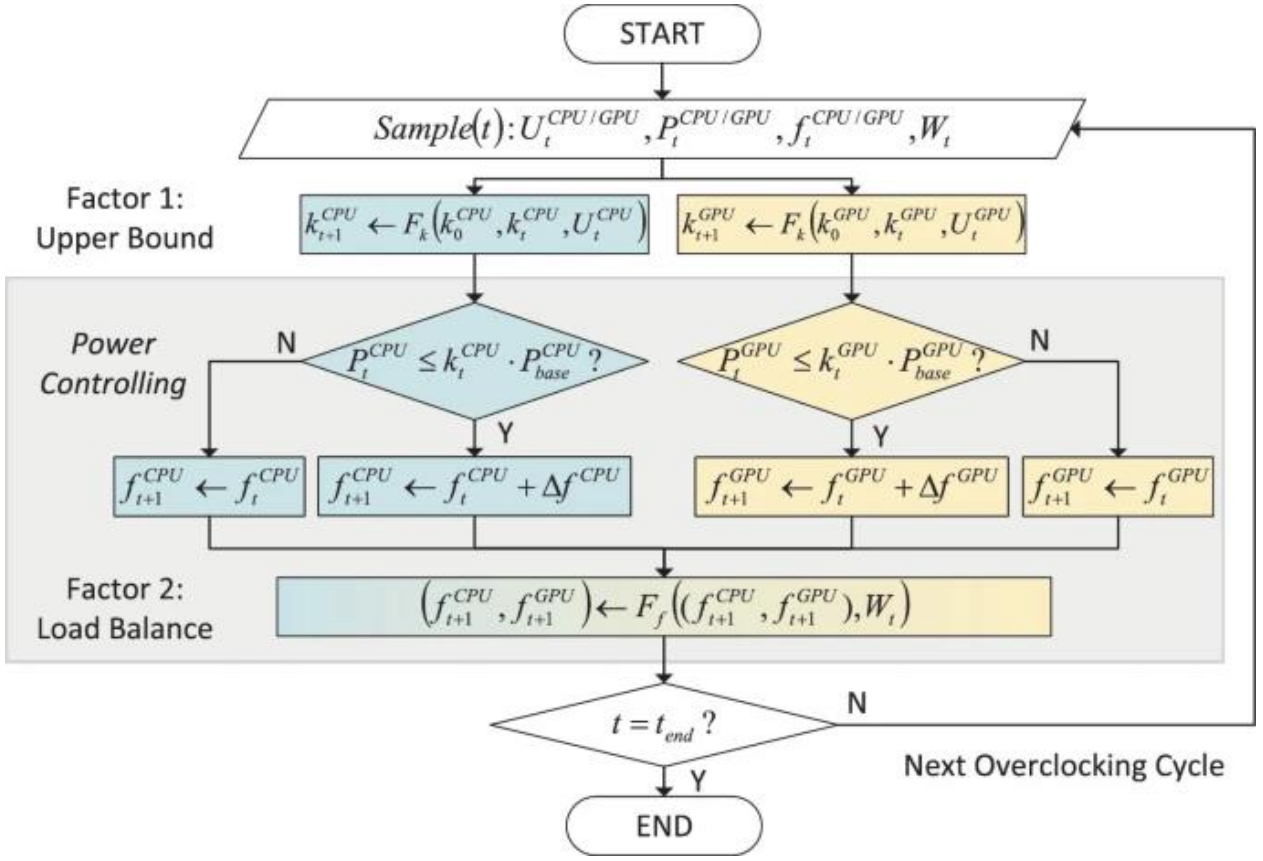
- What should be the upper power bound during overclocking?
  - Can the bound be dynamic instead of static?
- How can overclocking be coordinated between CPU & GPU?
  - Consider a load-imbalance factor?

Target: Constant total energy

- $P_{instantaneous} \not> P_{upper-bound}$
- $f \not> f_{upper-bound}$

22nd Int. Conf. on Algorithms and Architectures for Parallel Processing, Denmark, 2022, DOI: 10.1007/978-3-031-22677-9_14

Adaptive Overclocking Algorithm

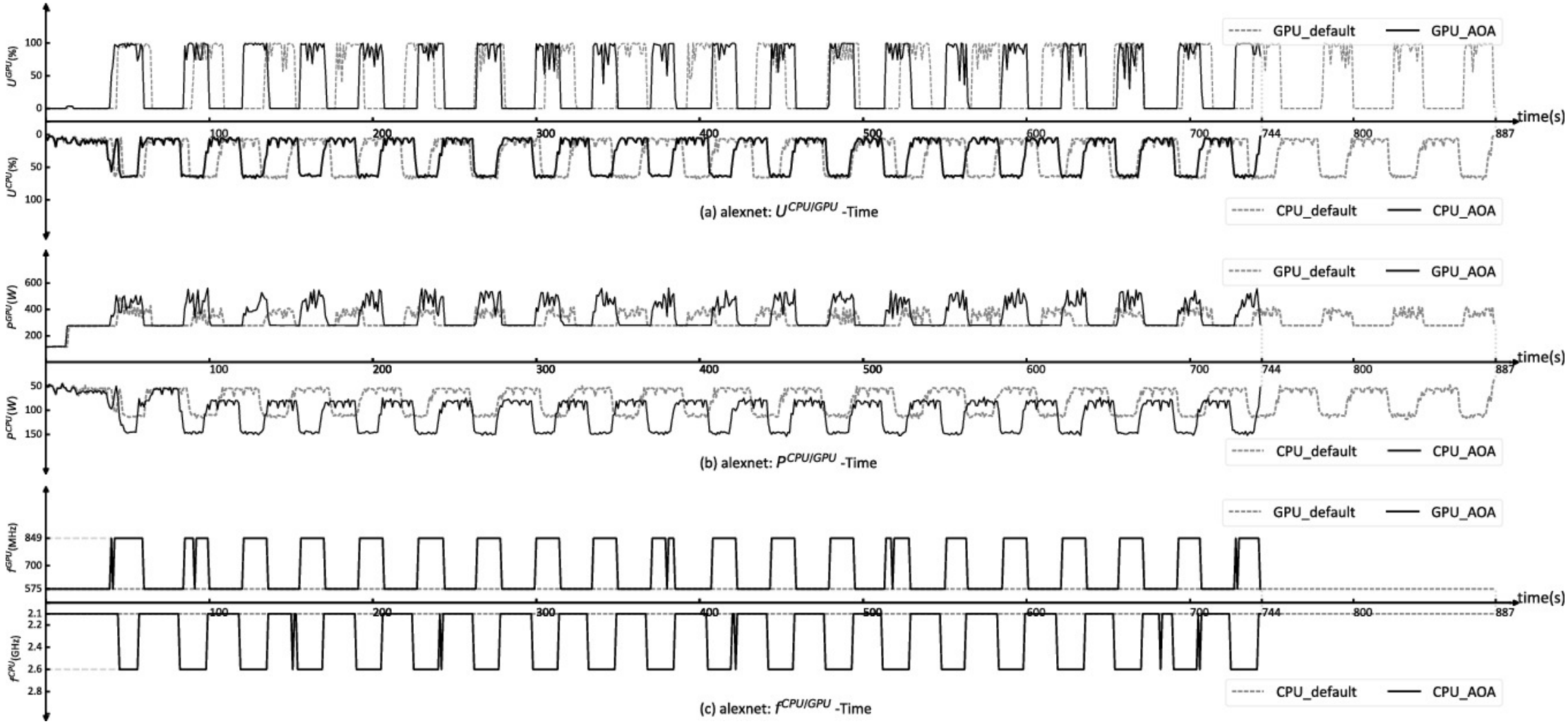


k = Power upper bound factor $(1 \le k \le 1.2)$

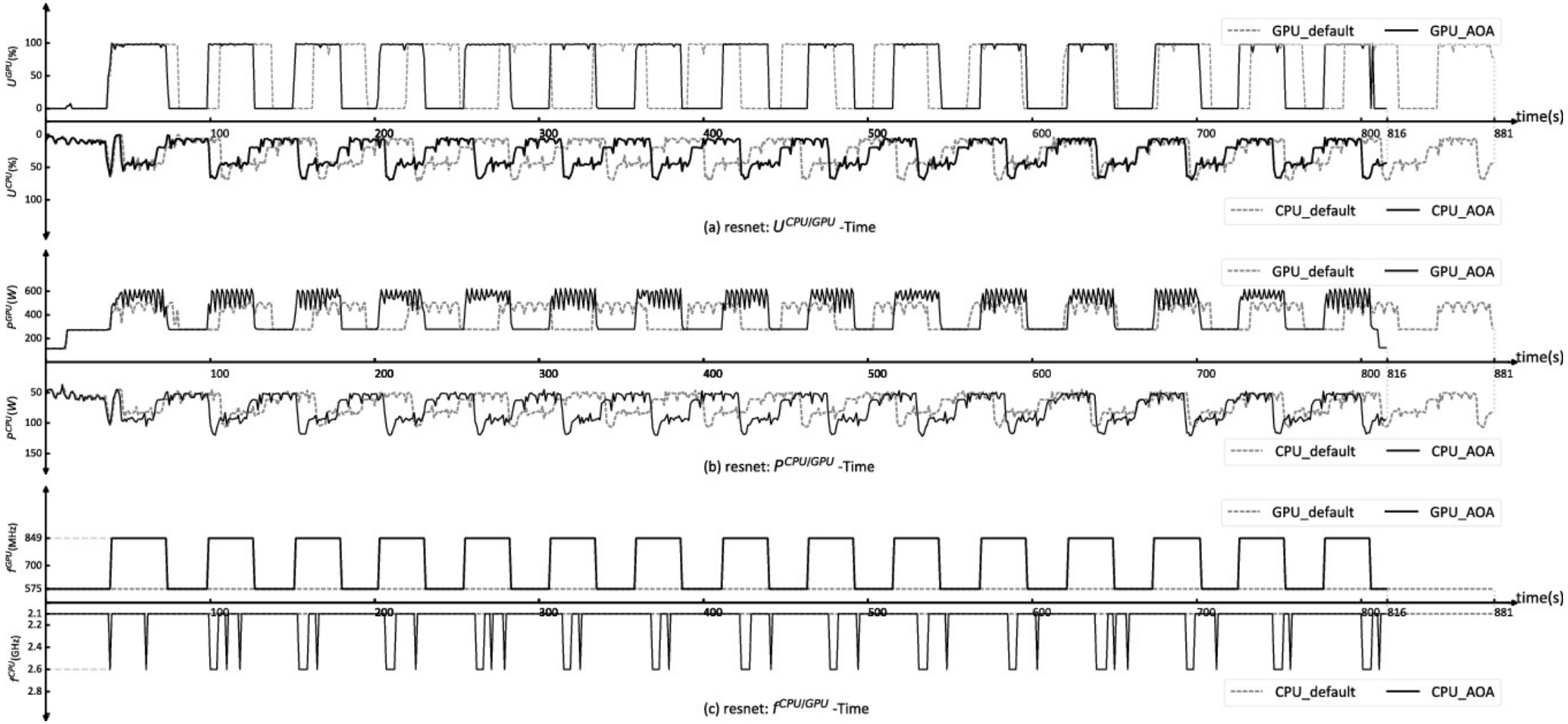
U_t = Utilization %

$W_t = \frac{U^{GPU} - U^{CPU}}{U^{GPU} + U^{CPU}}$ = Load imbalance factor

continued…

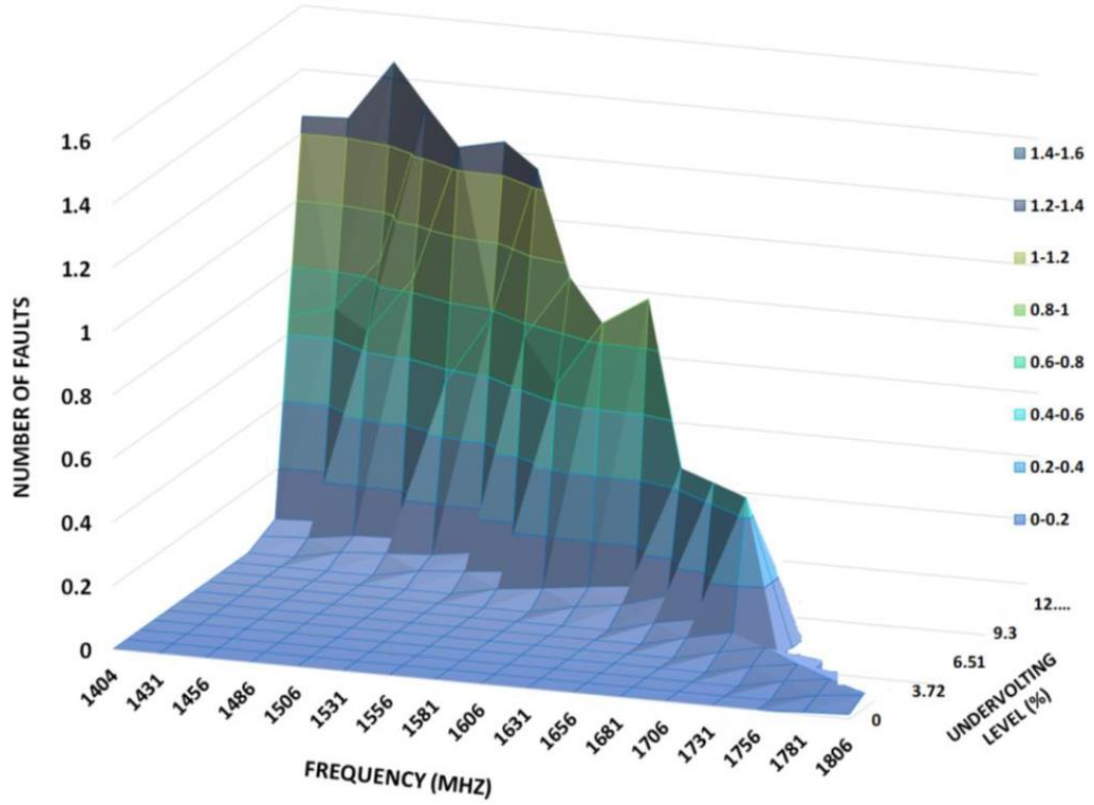

GPU_default
GPU_AOA
CPU_default
CPU_AOA
time(s)
100
200
300
400
500
600
700
744
800
887
(a) alexnet: $U^{CPU/GPU}$ -Time
(b) alexnet: $P^{CPU/GPU}$ -Time
(c) alexnet: $f^{CPU/GPU}$ -Time

continued…


GPU_default
GPU_AOA
CPU_default
CPU_AOA
time(s)
100
200
300
400
500
600
700
800
816
881
(a) resnet: $U^{CPU/GPU}$ -Time
(b) resnet: $P^{CPU/GPU}$ -Time
(c) resnet: $f^{CPU/GPU}$ -Time

GPU Undervolting

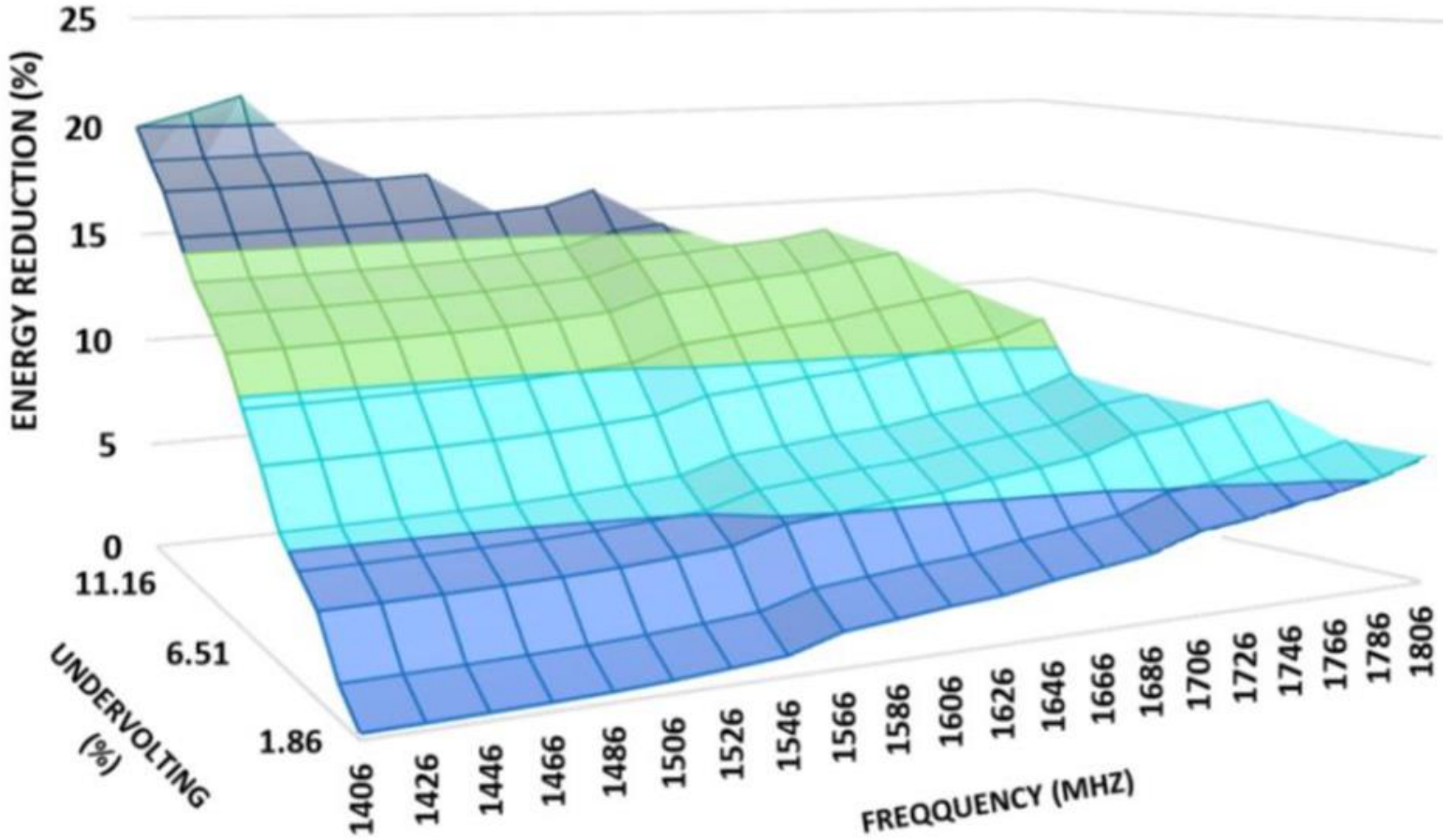
- There exists about 20% voltage guard-band on different GPU cards.
- Overclocking & Undervolting will increase error probability:
  - Silent Data Corruption
  - OS crash
  - Driver error
- Incremental checkpoint & recovery technique must be used.
  - Additional energy overhead



ACM/IEEE Int. Symp. on Low Power Electronics and Design, Boston, MA, 2020, DOI: 10.1145/3370748.3406553

continued…

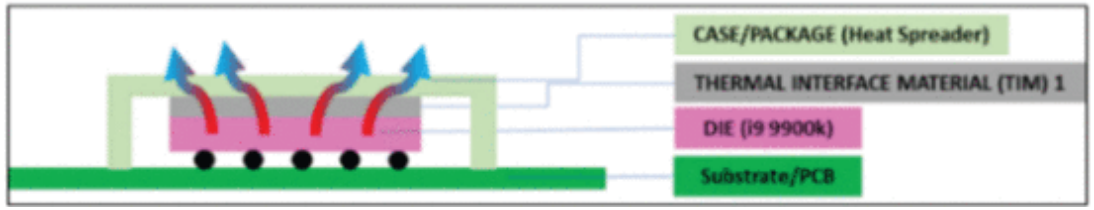
$10\text{k} \times 10\text{k}$ matrix multiplication
Platform: GeForce GTX 980



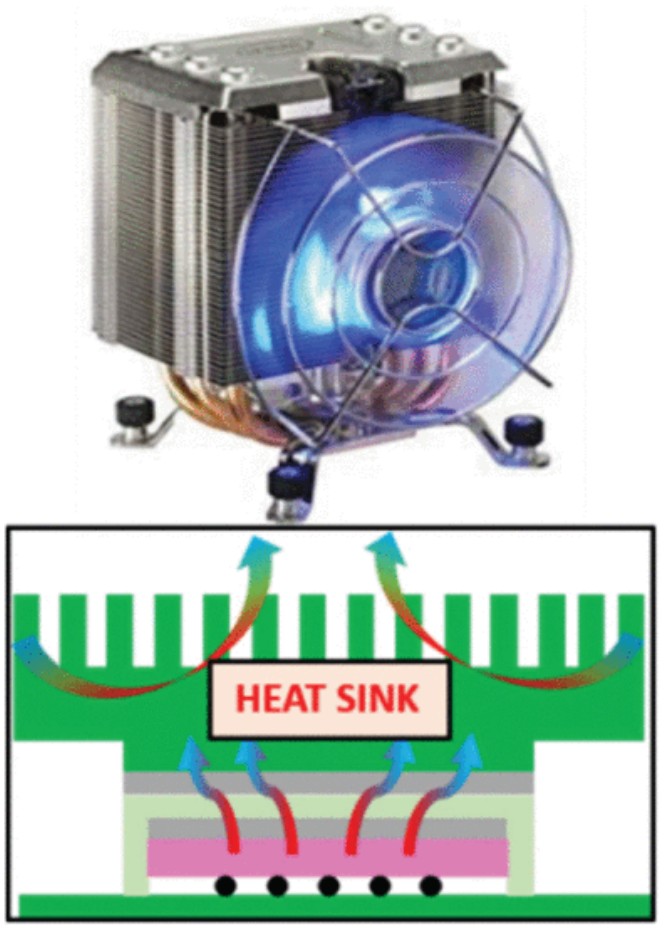
Cooling

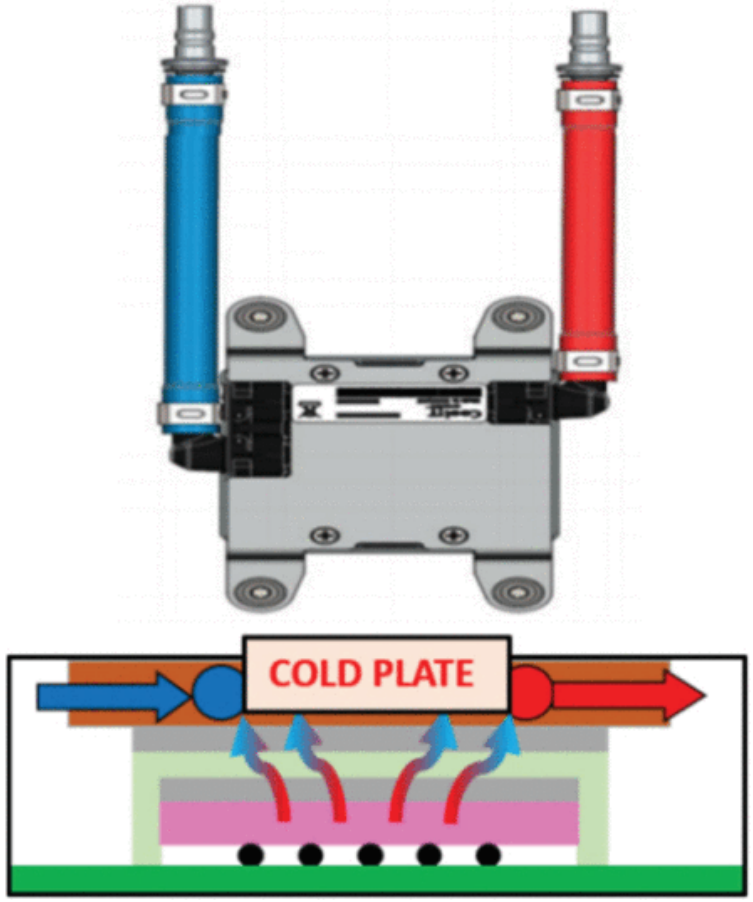
- During overclocking, power levels can go above TDP.
- T_j should not go beyond max rating.

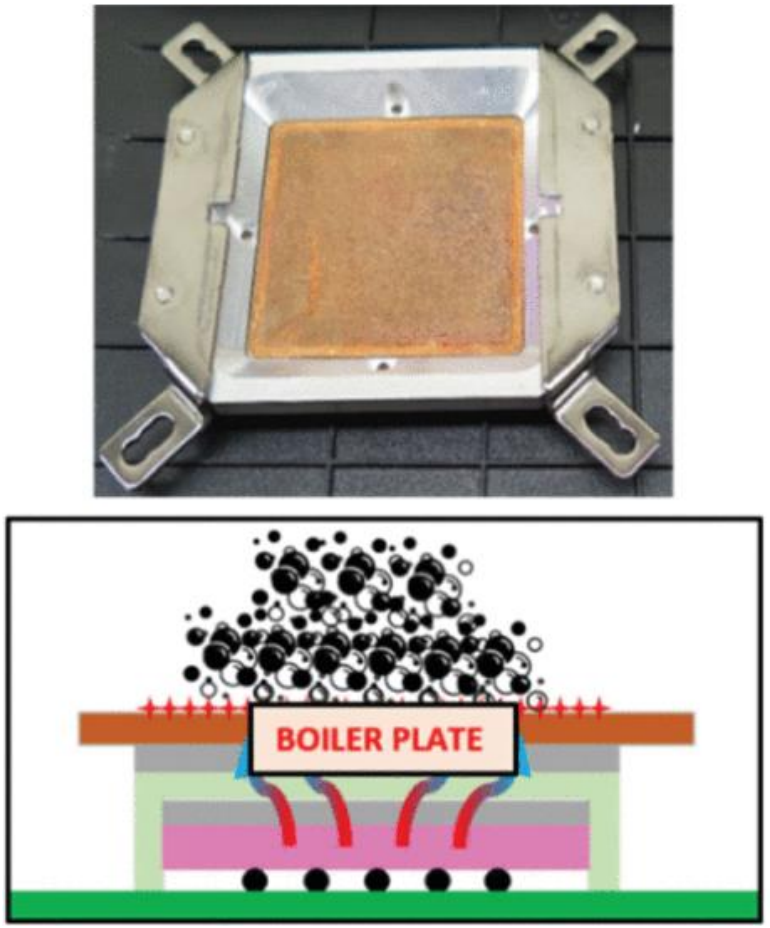
Die/Chip size	18mm x 9 mm
IHS/ Lid size	37.5 mm x 37.5 mm x 5 mm Material: Copper
TIM 1	Indium (k = 86 W/mK), Thickness = 0.2mm
TIM 2	Grease (k = 5 W/mK) Thickness = 0.1 mm



continued…

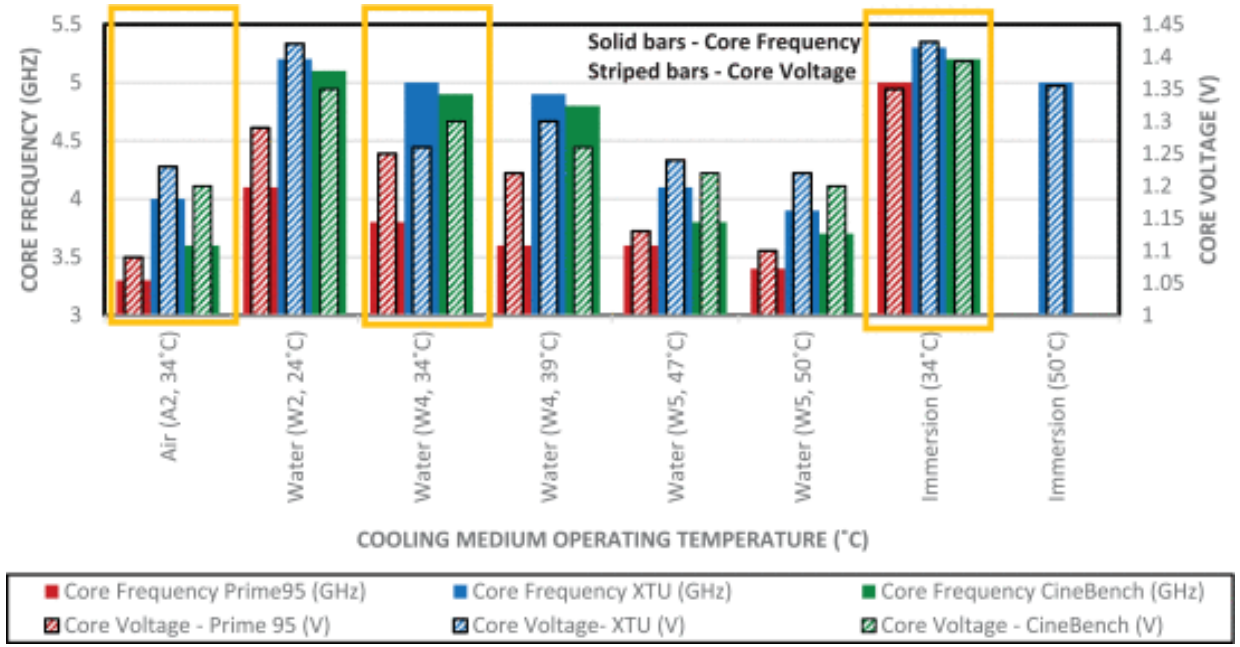


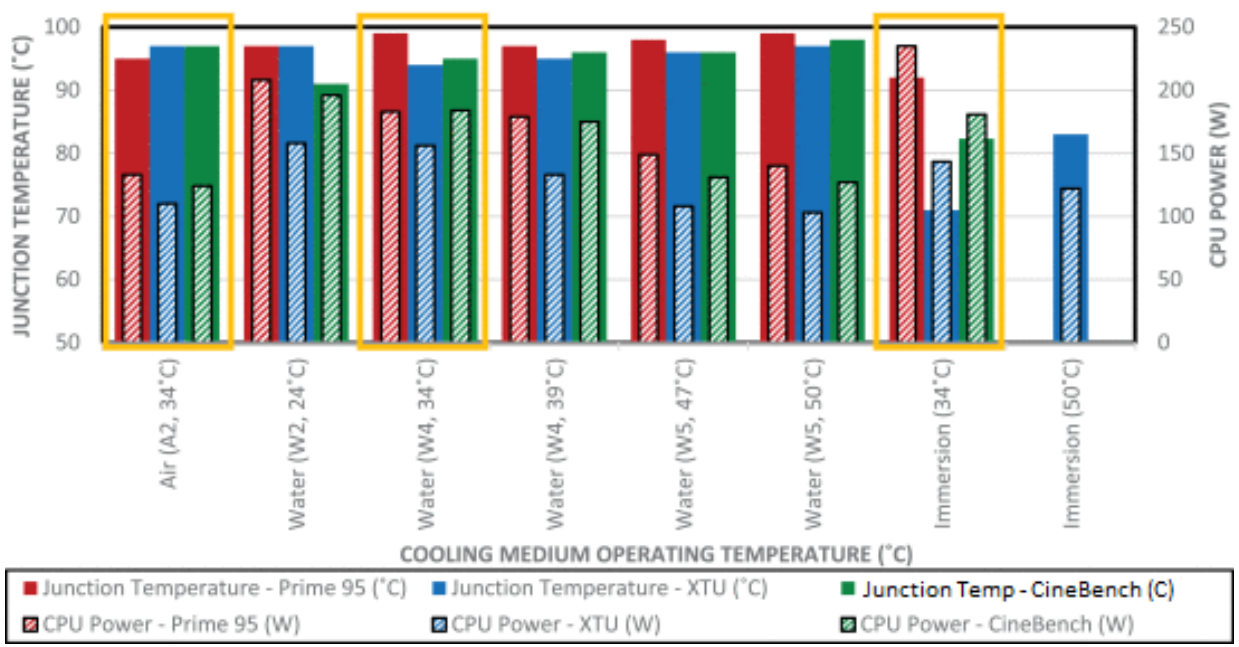


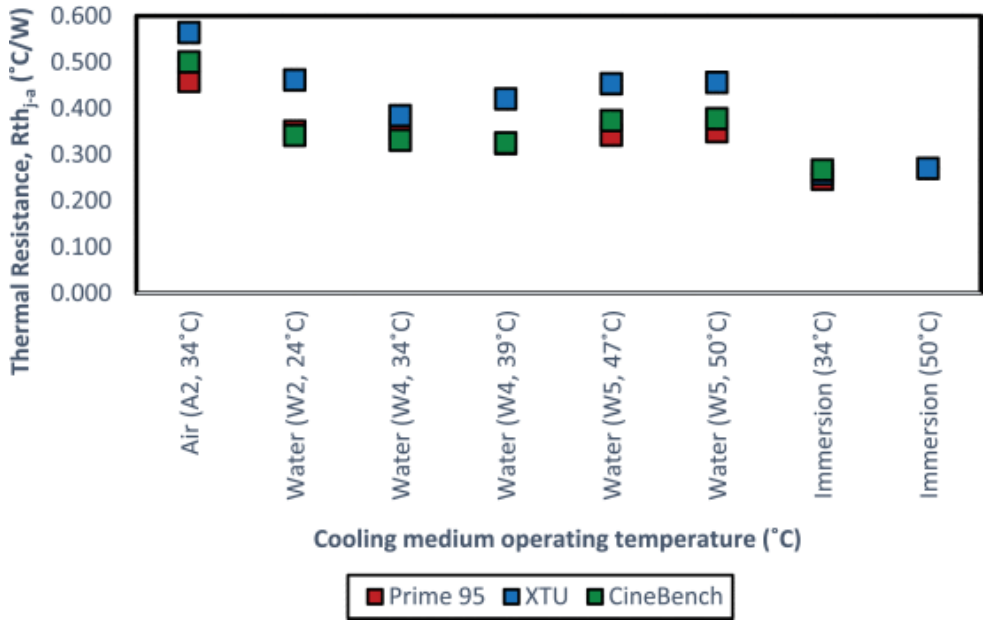


IEEE Transactions on Components, Packaging and Manufacturing Technology, 2021, DOI: 10.1109/TCPMT.2021.3106026

continued…



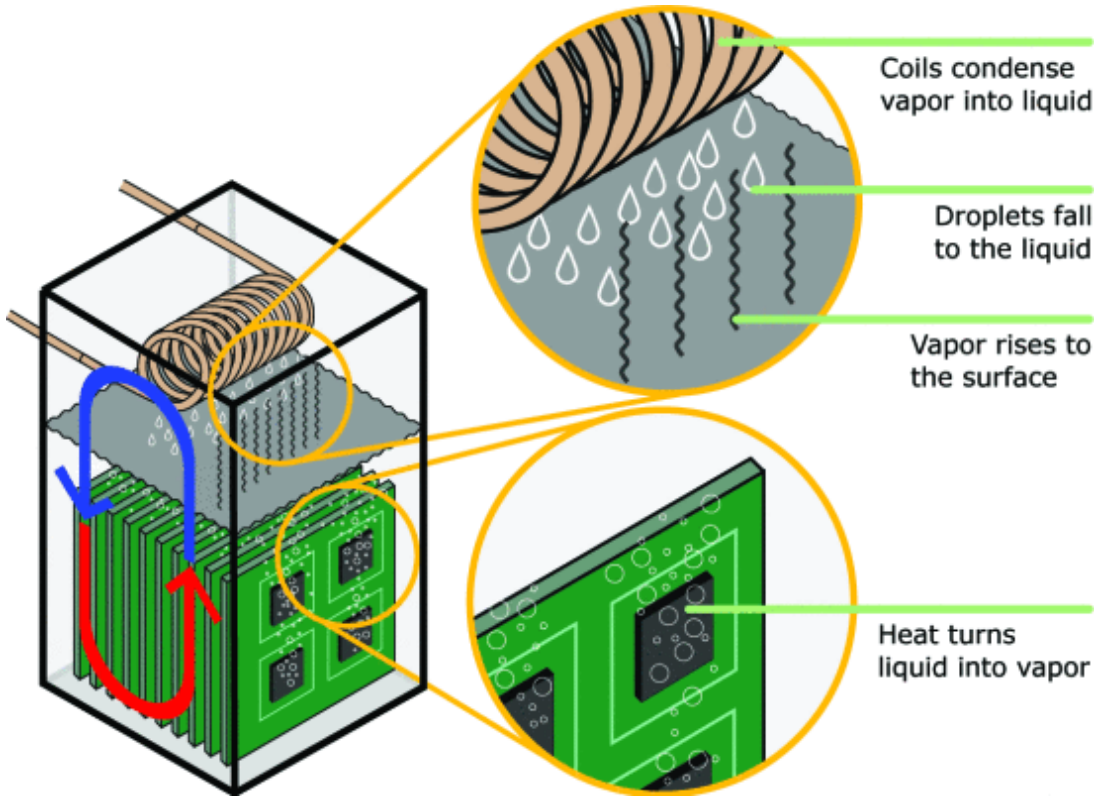




$$R_{thermal} = \frac{T_j - T_{amb}}{Q_{CPU}} (°C/W)$$

Cooling at Overclocked Datacenters

- Traditionally, air cooling has been employed.
  - ➢ Mechanical chiller, Water side economizer
- Air cooling won't be able to keep up with TDP.
- Liquid cooling and immersion cooling is now becoming the norm.

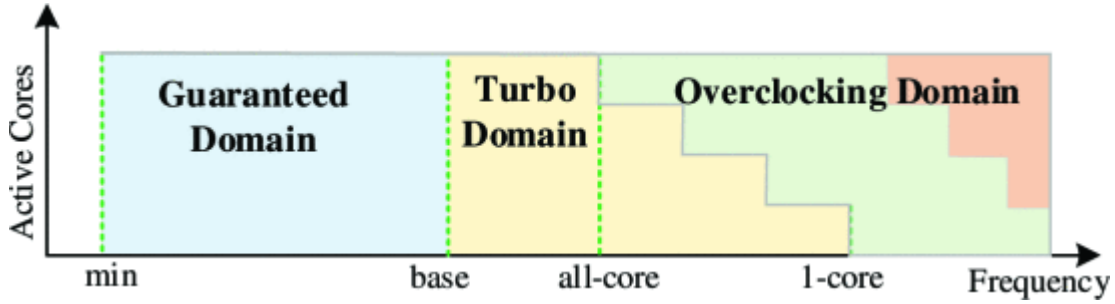


ACM/IEEE 48th Annu. Int. Symp. on Computer Architecture, Spain, 2021, DOI: 10.1109/ISCA52012.2021.00055

Benefits of Overclocked Datacenters

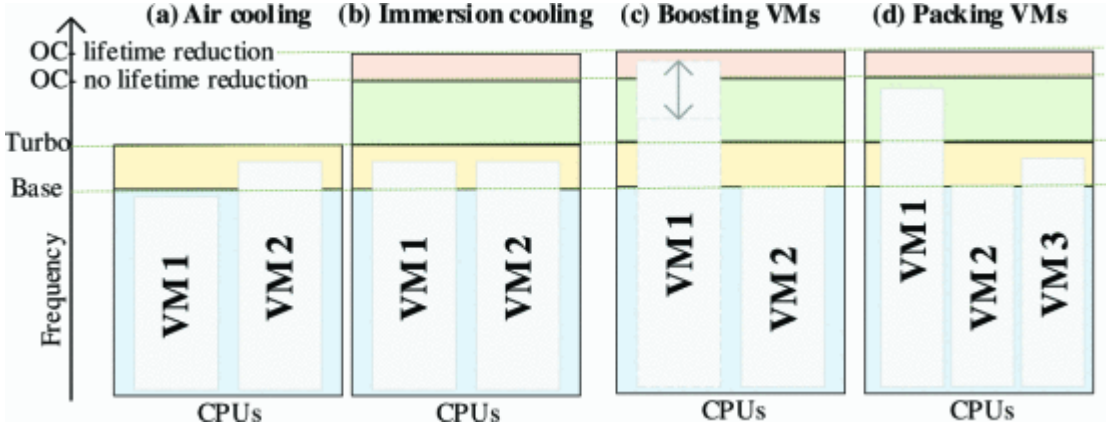
(1) High-performance VM

- Go beyond Turbo



(2) Dense VM packing

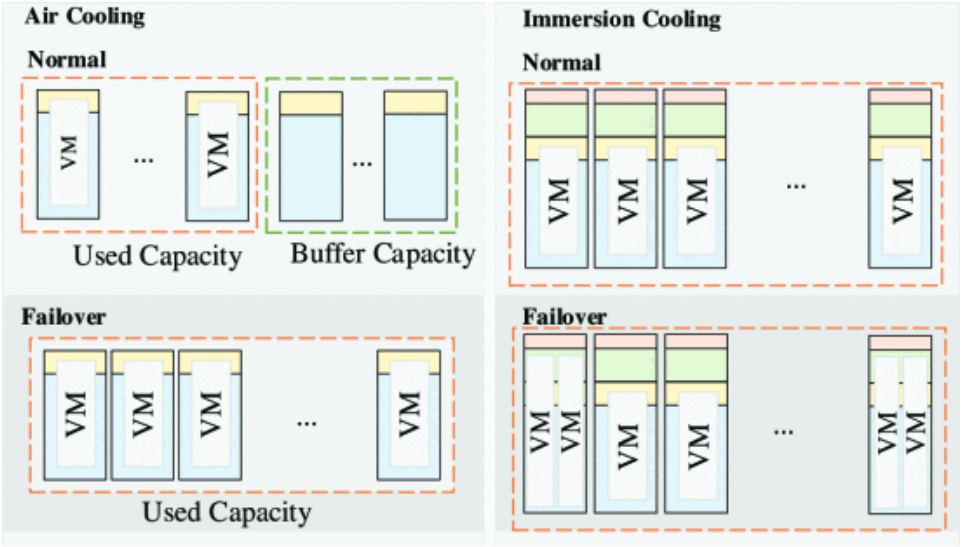
- Fight oversubscription



continued…

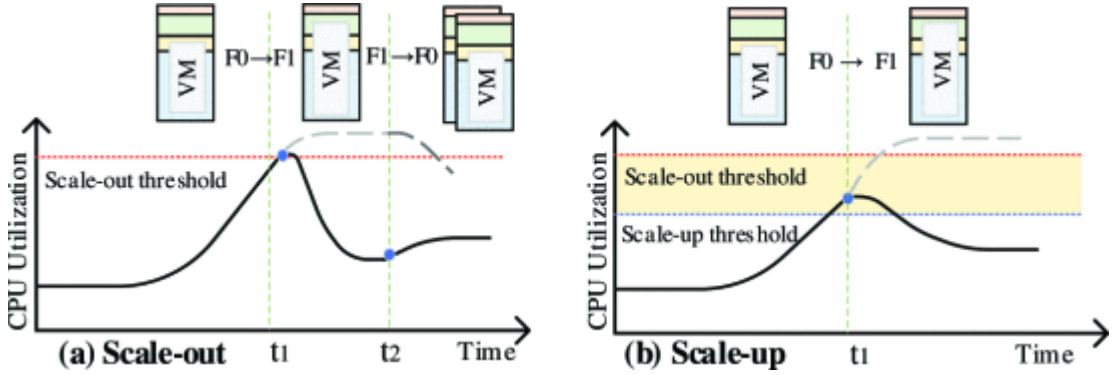
(3) Reserve Reduction

- Upon infrastructure failure, recreate affected VM and overclock



(4) Auto-scaling

- Boost existing VM while new VM is deployed (a)
- Prevent scale-out altogether (b)



Issues of Overclocked Datacenters

(1) Power

- Cannot overclock indiscriminately

(2) Lifetime

Failure Mode	Dependency T_j	Dependency ΔT	Dependency V	Description
Gate Oxide breakdown	✔	×	✔	A low impedance source to drain path
Electro-migration	✔	×	×	Material diffuses compromising gate structure
Thermal cycling	×	✔	×	Micro-cracks due to expansion-contraction

(3) Computational Stability

- Excessive overclocking (>23%) will affect stability.

continued…

(4) Environmental impact

- Overclocking at datacenters is a big contributor of CO_2.

(5) Cost of ownership

- Immersion cooling can provide up to 7% reduction in cost per physical core in comparison to air-cooled datacenters.

(6) Workload prediction

- Cloud providers have little or no knowledge of the workloads running inside the VMs.

Overclocking Smartphones

- Even under default settings, sustained performance-intensive workloads can trigger thermal throttling.
- THERMACLOCK
  - Estimate ambient temperature within 2°C
  - Profile workloads to obtain power estimates
  - Identify overclock-safe situations

  Evaluation: Three AI benchmarks (image classification, object detection, video upscaling)

Overclock-safe		*Overclock-safe* (predicted)	
		Yes	No
Overclock-safe	Yes	5,582	2,796
	No	1,835	25,675

Platform: Google Nexus 5

IEEE Int. Symp. on Performance Analysis of Syst. and Software, Stony Brook, NY, 2021, DOI: 10.1109/ISPASS51385.2021.00039

Conclusion

- The thermal challenge lies in managing heat flux (W/cm^2) rather than TDP (W).
- It is often cheaper (when all costs are considered) to buy faster hardware rather than overclocking an older component.
- Overclocking is worthwhile only if performance gains justify:
  - ✓increased cost of maintenance
  - ✓reduction in reliability and lifespan

Questions?
Comments?

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE