EPI TUTORIAL: POWER ASPECT

CO-LOCATED WITH THE HIPEAC2020

BOLOGNA, ITALY

22 JANUARY 2020

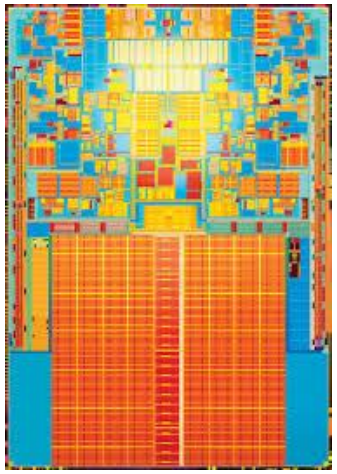
European Processor Initiative

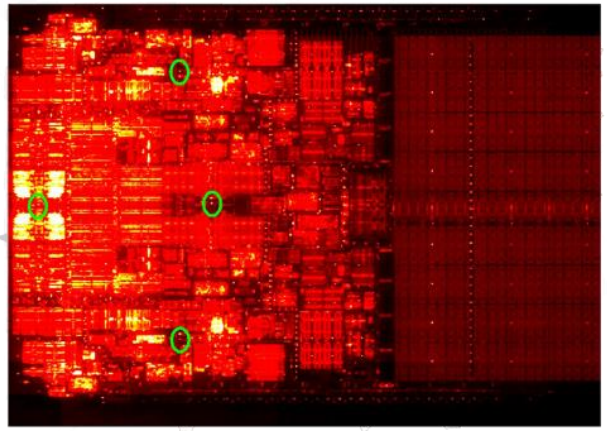
FRAMEWORK PARTNERSHIP AGREEMENT IN EUROPEAN LOW-POWER MICROPROCESSOR TECHNOLOGIES

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON 2020 RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 826647

European
Processor
Initiative

Intel Pentium - *1 core*





End of Dennard's Scaling
=> Power density increases
=> Thermal issues

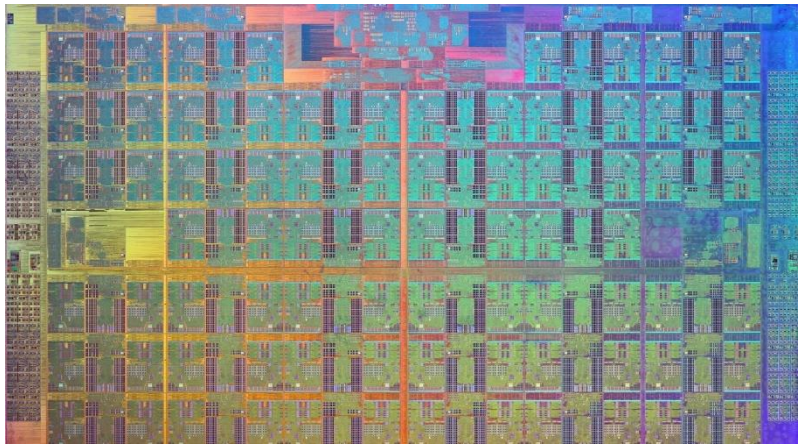
Same Power Budget !

1. ↓ frequency
2. ↓ # active core at the same time
3. ↑ cooling cost

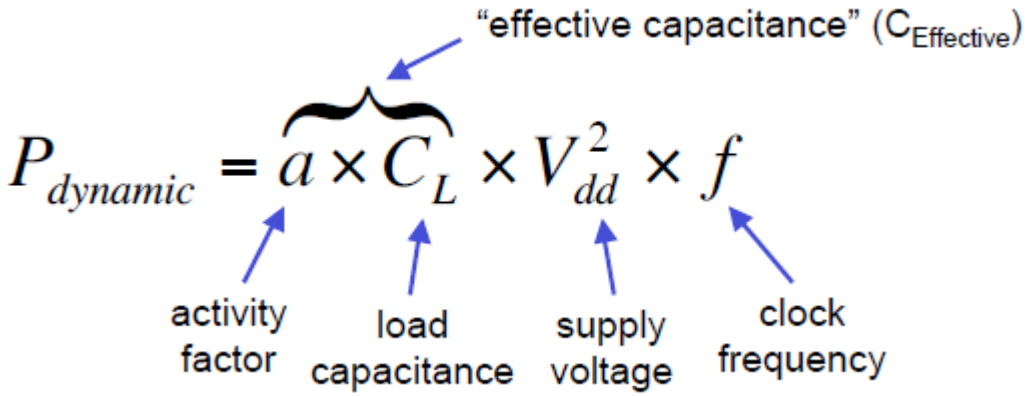
Moore's Law: "*The number of transistors and resistors on a chip doubles every 24 months*"

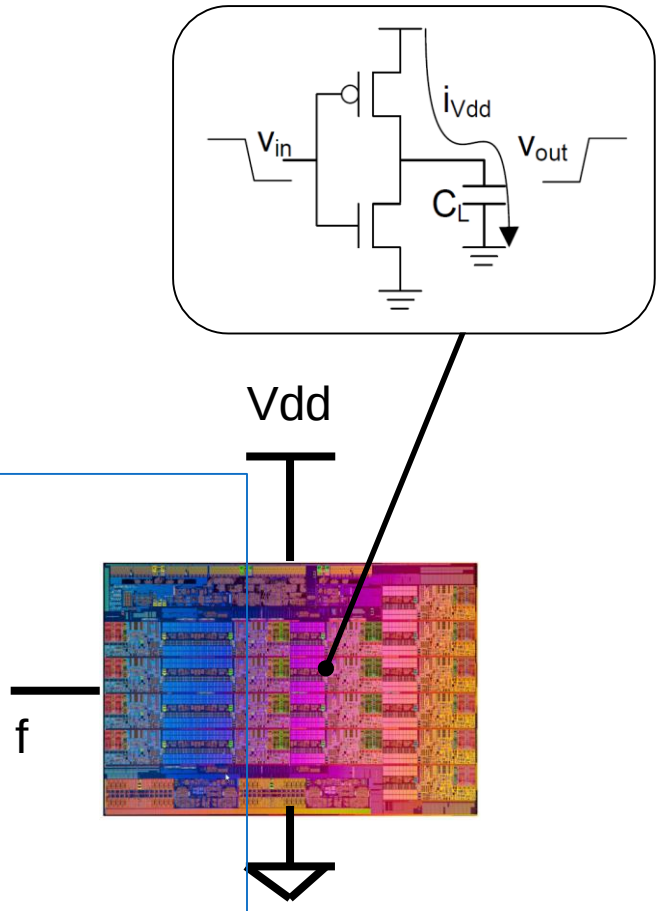
↑ theoretical max frequency
↑ #cores

Intel KNL – 72 cores



DYNAMIC POWER





- **Linear ↓ with ↓ $C_{Effective}$**
- **Linear ↓ with ↓ f**
- **Quadratic ↓ with ↓ V_{dd}**
- **Cubic ↓ with ↓ both V_{dd} and f**

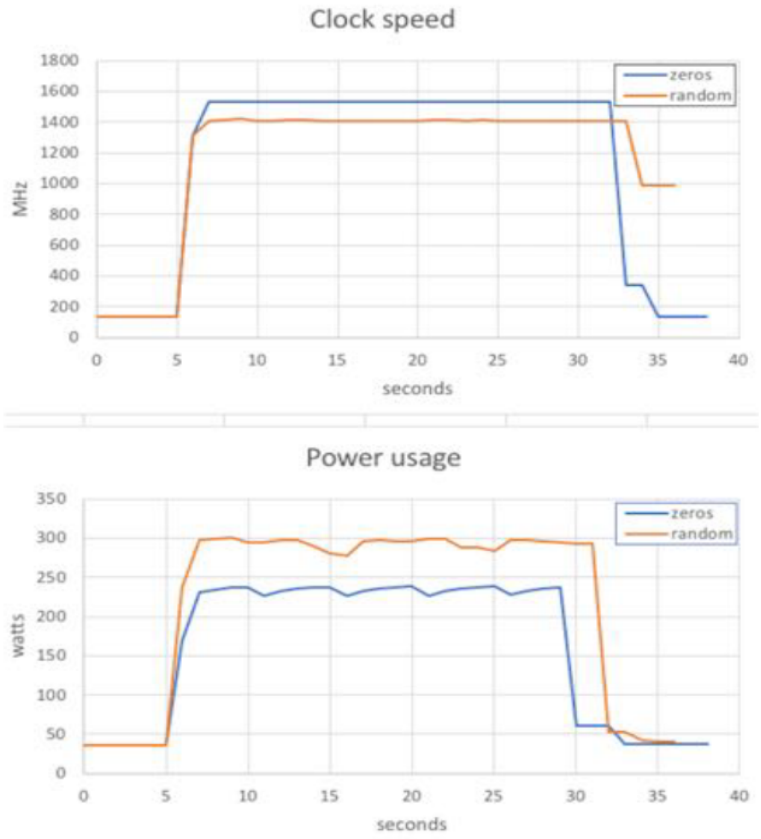
David H. Albonesi ACACES10

Ceff EFFECT

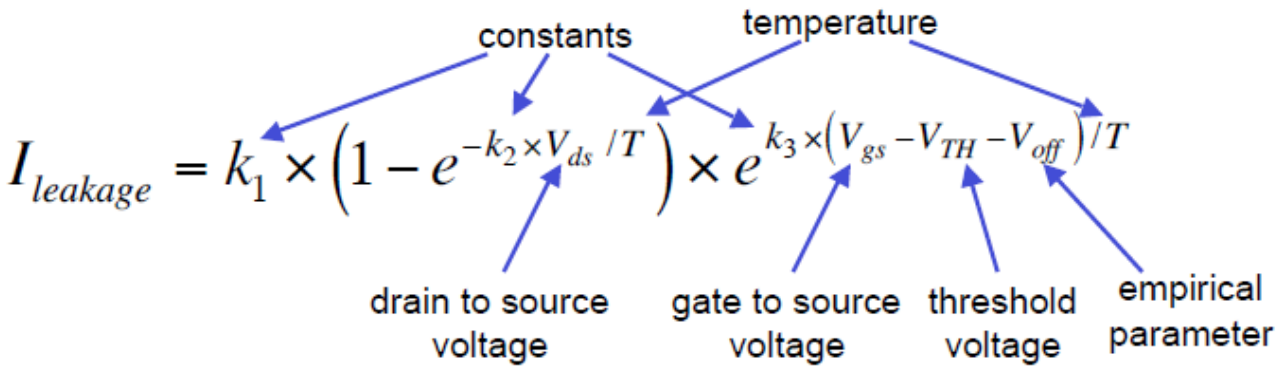
- TC/HGEMM has surprising data-dependent performance: **125 TF** theoretical peak, **113 TF** achievable on zero-filled matrices, **105 TF** peak on random CCC matrices, **~95 TF** peak on matrices with fully random FP16 entries

Issues

- Measurements on 1 Summit GPU using `nvidia-smi`
- Data-dependent performance of Tensor Cores is due to 300W power/frequency throttling of Voltas on Summit
- Baidu DeepBench GEMM benchmark has a bug (reported), incorrectly fills FP16 matrices with zeros instead of the intended random values, thus miscomputes GPU performance



SUB-THRESHOLD LEAKAGE CURRENT



- **Exponential ↓ with ↓ V_{gs} (~V_{dd})**
- **Exponential ↓ with ↑ V_{TH}**
- **Exponential ↓ with ↓ T**

David H. Albonesi ACACES10

LEAKAGE IN SUPERCOMPUTING

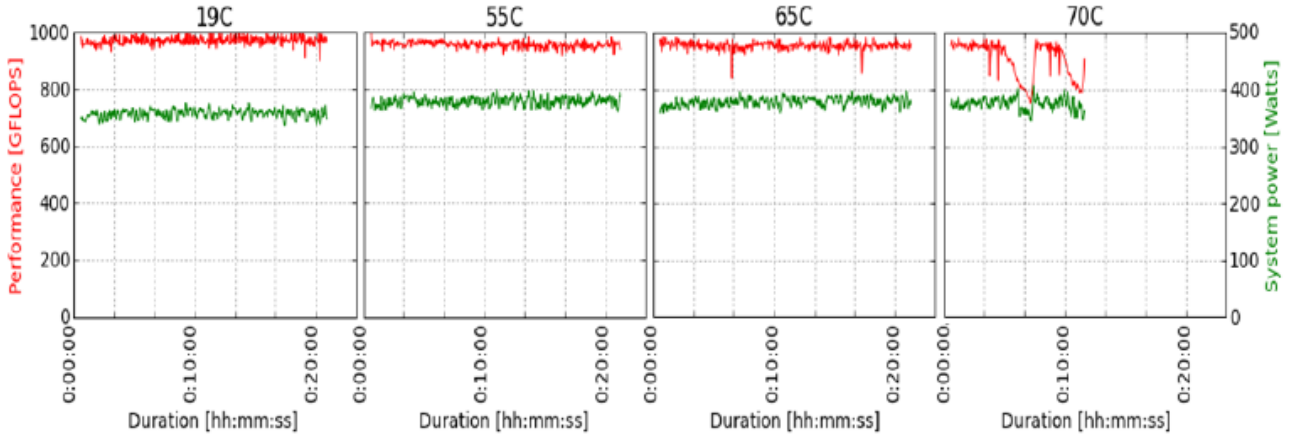


Figure 3. Illustration of system performance and power profiles at T_{LC} = 19, 55, 65 and 70 °C

A. Moskovsky et al., "Server level liquid cooling: Do higher system temperatures improve energy efficiency?" Supercomputing frontiers and innovations, 3(1):67-74, 2016.

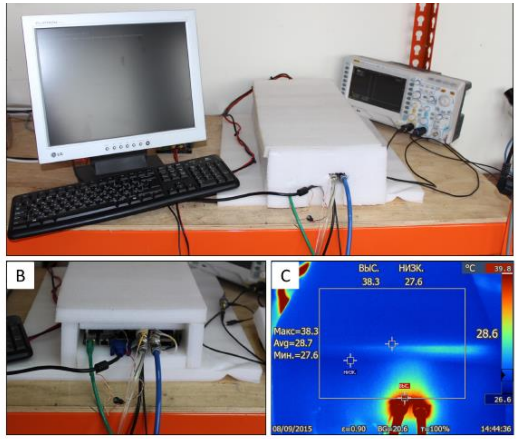


Figure 1. a) Overview of the experimental setup; b) liquid cooling and system connection Ti32 view (T_{max} = 38.3 °C, T_{min} = 27.6 °C)

Table 1: Performance and power characteristics at different coolant temperatures

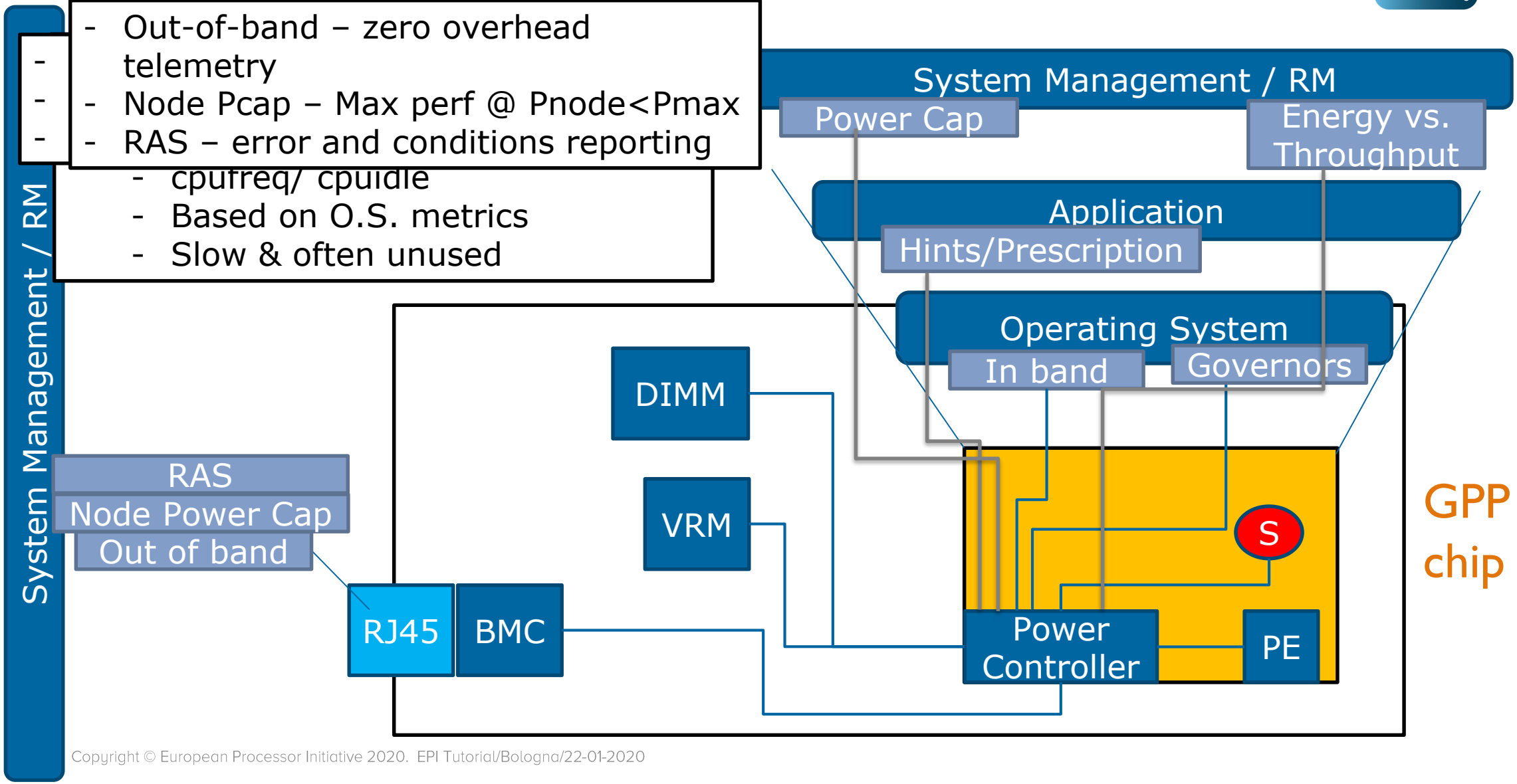
T_{LC} (°C)	Performance (GFLOPS)	System power (W)	Efficiency (GFLOPS/W)	Efficiency decrease (%)	Possible energy reuse[1](%)
19	974	359	2.72	–	-1.8
45	985	388	2.54	7.0	7.8
50	985	390	2.53	7.5	9.7
55	981	395	2.49	9.4	10.8
60	976	396	2.46	10.3	12.3
65	969	398	2.44	11.5	14.9

[1] assuming environment temperature T_0 = 25 °C

WP3 – POWER MANAGEMENT & CONTROLLER

- Out-of-band – zero overhead telemetry
- Node Pcap – Max perf @ Pnode<Pmax
- RAS – error and conditions reporting

- cpufreq/ cpuidle
- Based on O.S. metrics
- Slow & often unused



POWER MANAGEMENT SOA & REQUIREMENTS

	Intel	IBM	ARM	AMD	Cray	Fujitsu
Monitor (Domain,Granularity)	S, M, A, T **1ms**	N, S, M, A, T, U **500us**,10ms aggregation 16ms for T & U, 100ms aggregation	S, M, T 1-10KHz with SCP	N, S, M, A, T 1 sec (C), 1ms (G)	N, S, M, A, N OOB (100ms)	N, S, C, M **1ms** (N), ~ns - model based (C)
Control (Domain,Granularity)	S, M RAPL **1ms** (in-band), DVFS **500us**	N, S, M, A 10-100ms	S, M 1-10KHz (100ms to 1s)	N, S, M, A ~secs	N, S, M, A DVFS, RAPL, min-max range, 10-30s at job launch	S, C, M, **DVFS**, Decode Width, HBM2 B/W
Interfaces, Tools, etc	RAPL MSRS, msr-safe, libmsr, PAPI, likwid	**OpenBMC**, amester, **Memory Map**	ACPI, SCP (sys ctrl proc), IPA (intelligent allocator), PAPI	Likwid, PAPI, **Memory Map**	CapMC, PAPI, Cray **BMC interfaces**	Power API, PAPI

Source PowerStack19

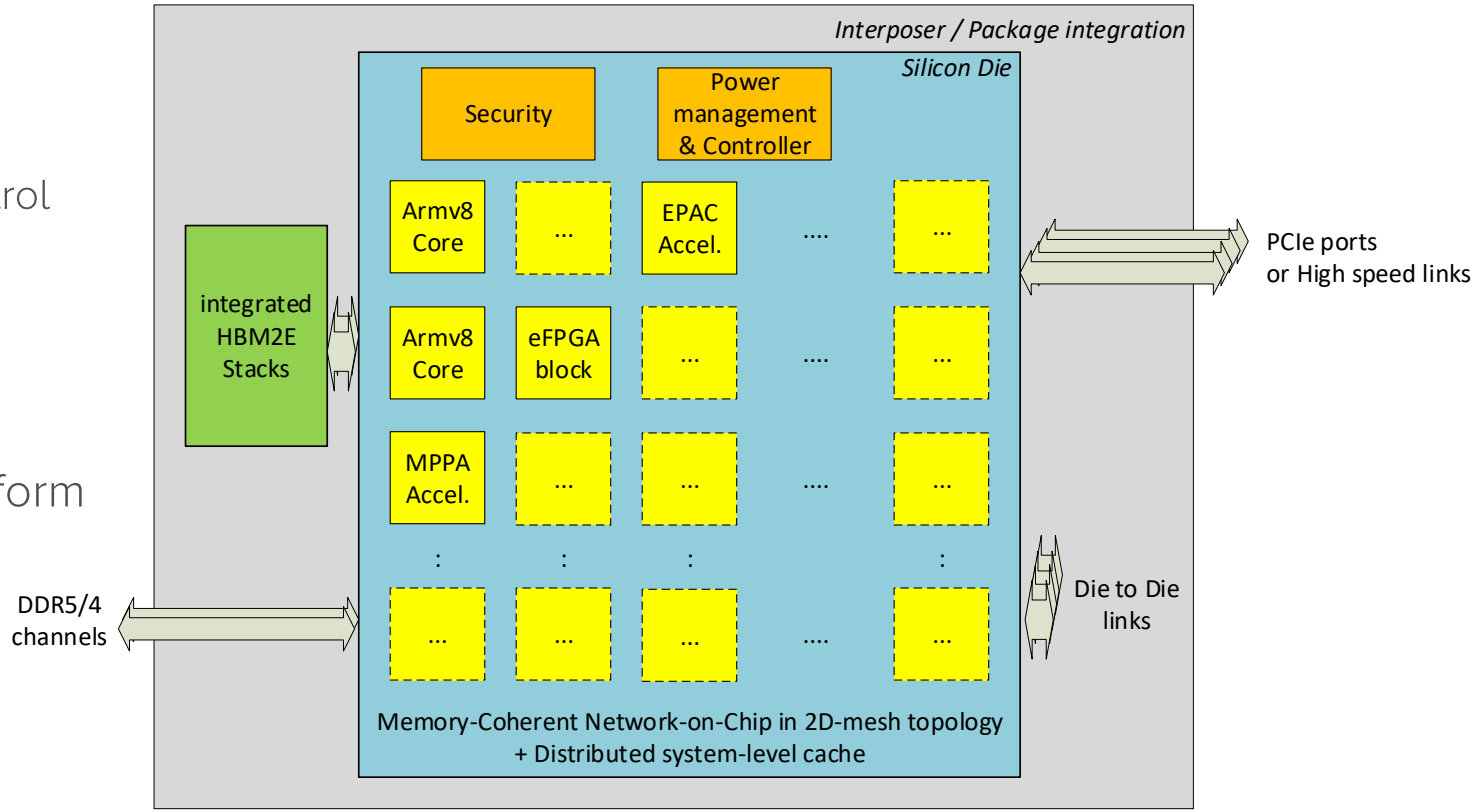
Socket (S), Core (C), Memory (M), Accelerator (G), Node (N), Utilization (U), Temperature (T)

EPI power management design is powered UNIBO and targets:

- Support for fine grain power monitoring, and control
- An higher performance power controller capable of supporting advanced power control algorithms.

GENERAL ARCHITECTURE

- Top level infrastructures
  - Power management & controller
  - Dedicated power management and control network
  - Security
- EPI Power Management Subsystem
- RISC-V ISA, Derived from the PULP platform
- Parallel processor w. DSP extensions
- Open-Source Design



THE POWER CONTROLLER FIRMWARE

T control

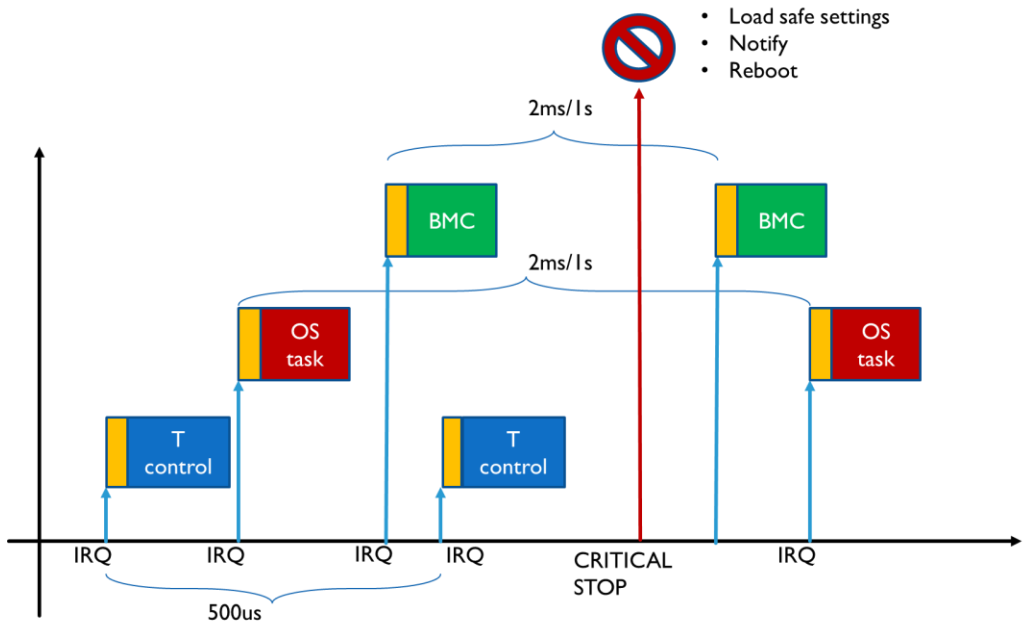
Watchdog reset
Write power controller settings
Write to internal memory telemetry data
Read PVT sensors
Read workload from O.S.
Read **target P/C state settings, power budget**
Read Pending BMC requests
Compute controller settings

PM task

- Read voltage regulator, power, status (VR)
- Power model update

BMC

- Read pending command queue
- Decode Command/data
- Perform action:
  - Change **target P/C state, power budget**
    - Set pending BMC
  - Ask telemetry data



CONCLUSIONS

- Power management is a key aspect of HPC processors
- Implemented by mean of a embedded computing subsystem with extensions for interfacing with the power management IPs.
- EPI will leverage a best-in class power management subsystem based on parallel architecture with DSP extensions.

