TECHNISCHE UNIVERSITÄT DRESDEN

Wake-up Latencies for Processor Idle States on Current x86 Processors

5th International Conference on Energy-Aware High Performance Computing (EnA-HPC)

Robert Schöne (robert.schoene@tu-dresden.de)
Daniel Molka (daniel.molka@tu-dresden.de)
Michael Werner (michael.werner3@tu-dresden.de)

ZIH
Center for Information Services & High Performance Computing

Overview

SPONSORED BY THE
Federal Ministry of Education and Research

- Introduction on processor idle states
  - Processor idle states in theory
  - Processor idle states in the field
- Why should you care?
- Measurement methodology
  - Instrumented kernel functions
  - Wake-up scenarios
- Results
- Summary

Introduction on Processor Idle States

SPONSORED BY THE
Federal Ministry of Education and Research

DVFS

$$P = \underbrace{\alpha\, C V^2 f}_{\text{Dynamic part}} + \underbrace{I_{static} V}_{\text{Static part}}$$

Introduction on Processor Idle States

SPONSORED BY THE
Federal Ministry of Education and Research

Clock gating

$$P = \underbrace{\alpha\, CV^2 f}_{\text{Dynamic part}} + \underbrace{I_{static} V}_{\text{Static part}}$$

Introduction on Processor Idle States

SPONSORED BY THE
Federal Ministry of Education and Research

Power gating

$$P = \underbrace{\alpha\, C V^2 f}_{\text{Dynamic part}} + \underbrace{I_{static} V}_{\text{Static part}}$$

Introduction on Processor Idle States -Theory

SPONSORED BY THE
Federal Ministry of Education and Research

- ACPI standard
- C0: The processor is executing instructions, P-States
- C1: Halt state
  - Return to C0 immediately
- C2
  - Return to C0 with delay
  - Processor responds to cache coherence traffic
- C3+:
  - Return to C0 with significant delay
  - Processor does not respond to cache coherence traffic
- Delays are handed over to OS via ACPI

Introduction on Processor Idle States – Intel

SPONSORED BY THE
Federal Ministry of Education and Research

C state	Core	Package
C0	Processor is actively executing instructions, P-States	
C1	Processor is inactive	If C1E is active: increase P-State to maximum
C2		Handle traffic from QPI / PCIe
C3	Flush caches to L3 cache, Clock gating	Disable ring, thus L3 cache inaccessible, L3 retains context Disable QPI / PCIe if latency allows it, DRAM self-refresh
C6	Save architectural state to SRAM, Power gate	
C7		*Flush L3, power gate L3 and SA*

Introduction on Processor Idle States – AMD Family 15h

SPONSORED BY THE
Federal Ministry of Education and Research

C state	Module	Package
C0	Processor is actively, executing instructions P-States	Northbridge P-States, *Memory P-States*
Cx (up to 3, programmed by BIOS)	**Flush L1 and L2 cache if timer expires,** **Clock gate module,** Store architectural state in DRAM, Power gate module, Pop Down P-State	*DRAM self-refresh,* *Northbridge clock and power gating,* *Package power off*

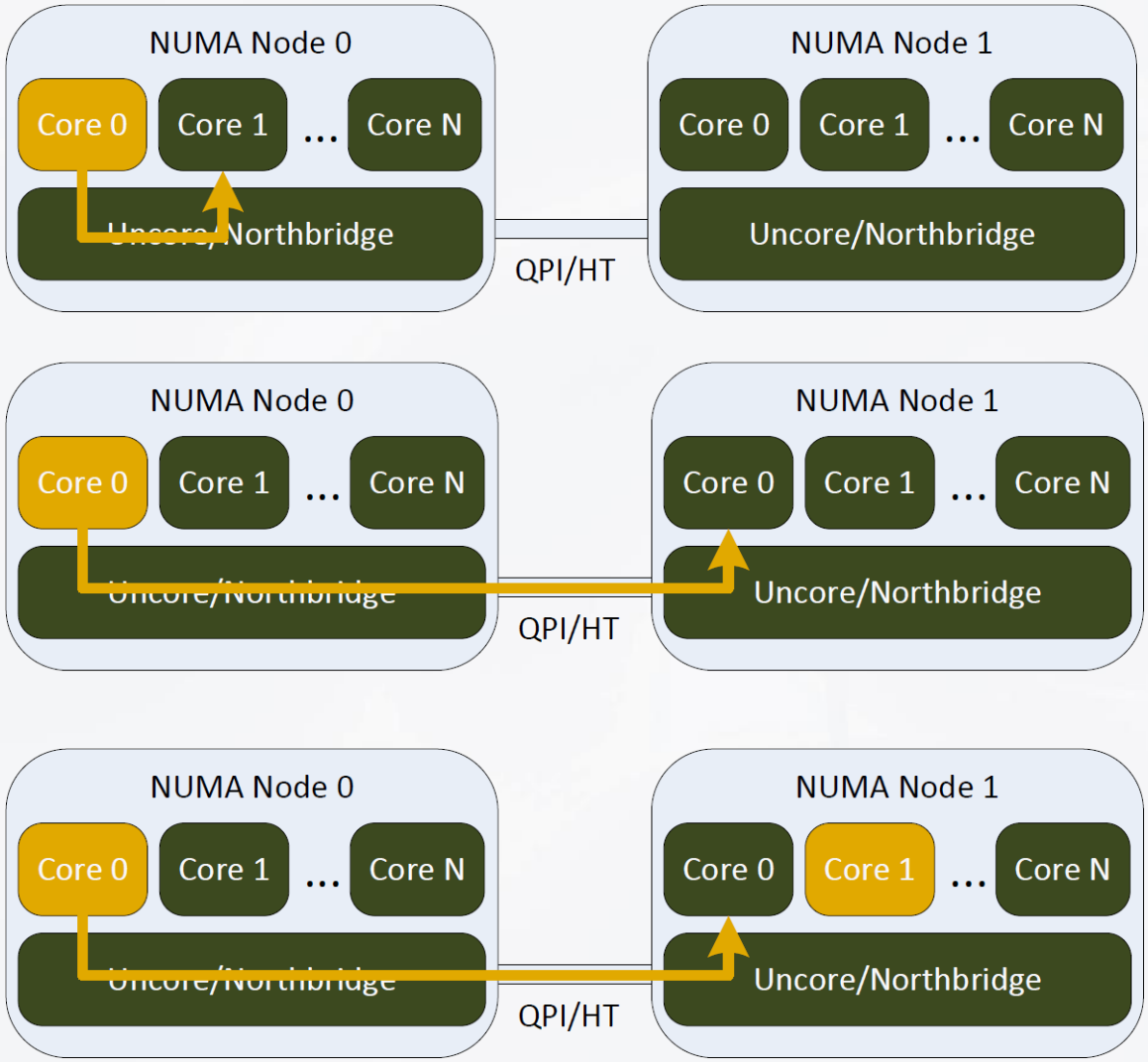
Why Should You Care?

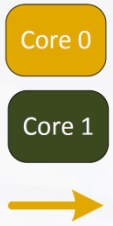
SPONSORED BY THE
Federal Ministry of Education and Research

- Energy saving vs. responsiveness
- What if the latency numbers provided by the processor vendors are too high?
  - Use lower C States
  - Burn energy unnecessarily
- What if the latency numbers provided by the processor vendors are too low?
  - Use higher C states
  - Responsiveness and performance degrades
- Idle states might be used in the following cases:
  - OpenMP synchronization, blocking I/O, blocking MPI, Dynamic Concurrency Throttling

Measurement Methodology

SPONSORED BY THE
Federal Ministry of Education and Research





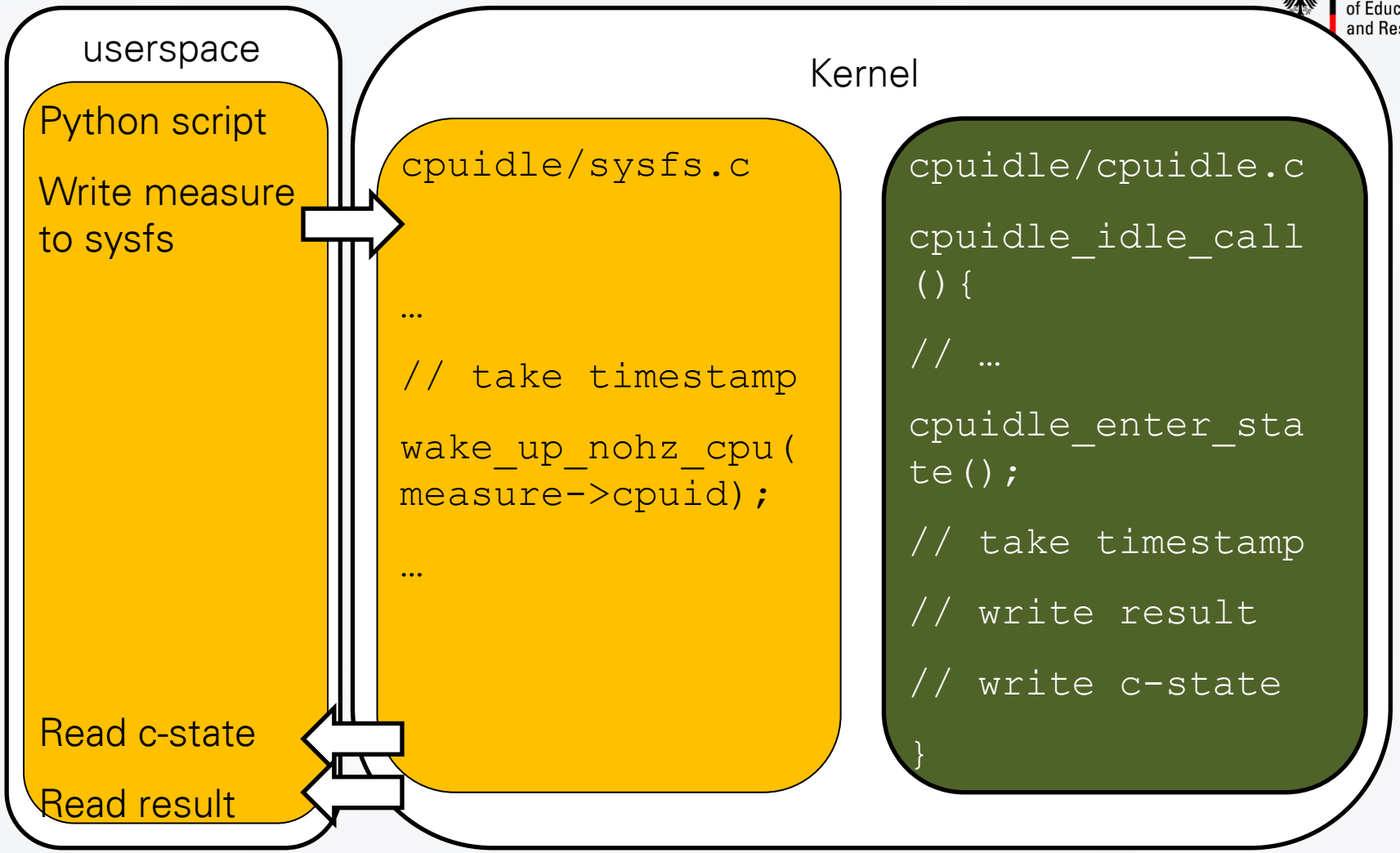
Local measurement

Remote idle measurement

Remote active measurement

Measurement Methodology

SPONSORED BY THE
Federal Ministry of Education and Research



Best case assumption with OS overhead, 400+ tests

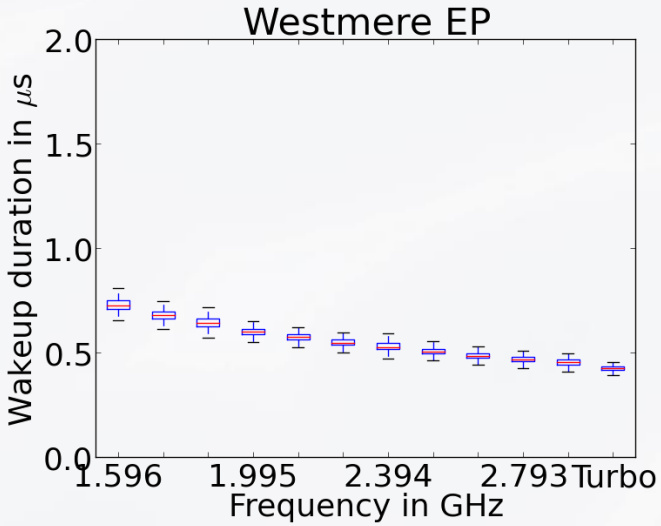
Test Systems

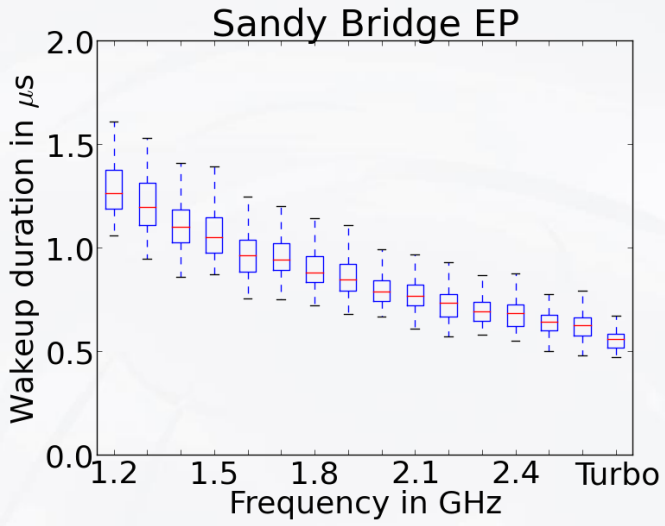
SPONSORED BY THE
Federal Ministry of Education and Research

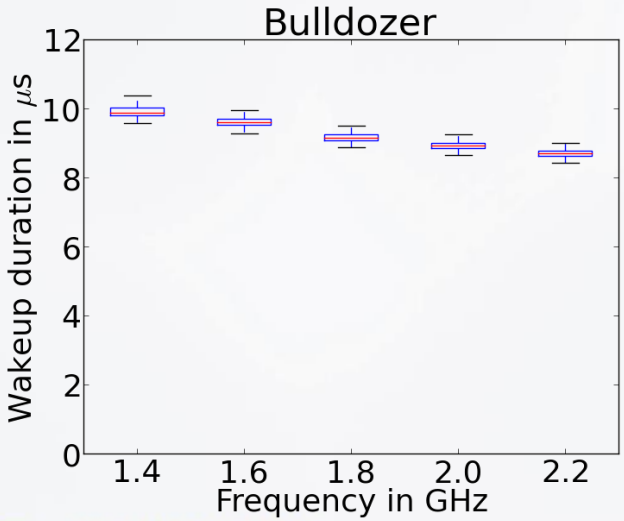
Vendor	Intel		AMD
Processor	Xeon X5670	Xeon E5-2670	Opteron 6274
Codename	Westmere-EP	Sandy Bridge-EP	Bulldozer
Cores	2x6	2x8	4x16
Base clock	2.933 GHz	2.6 GHz	2.2 GHz
Max Turbo Clock	3.333 GHz	3.3 GHz	3.1 GHz
Uncore/NB clock	2.666 GHz	-	2.0 GHz
C-States	C1, C3, C6	C1, C3, C6, C7	CC1, CC6
PC-States	PC1E, PC3, PC6		*n/a*

Results C1 (Local, According to ACPI: 3/2/0 µs)

SPONSORED BY THE
Federal Ministry of Education and Research



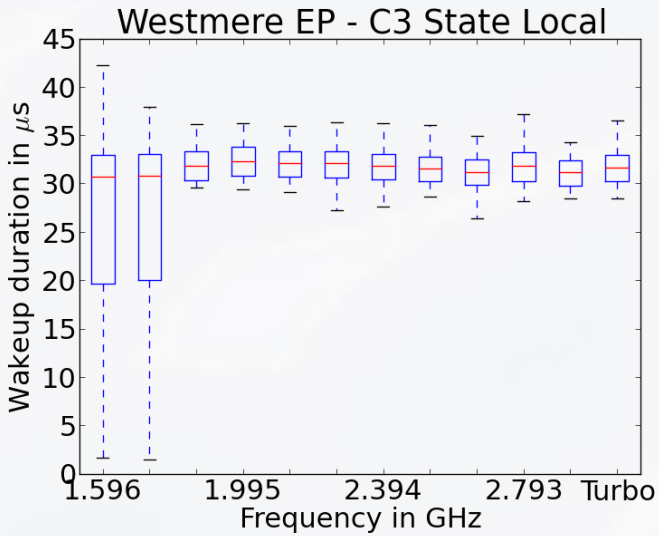




- Higher latency on newer Intel system
- AMD Bulldozer latency much higher than Intel latency
- Remote case increases latency by approx. 0.2 - 0.5 µs (not depicted)

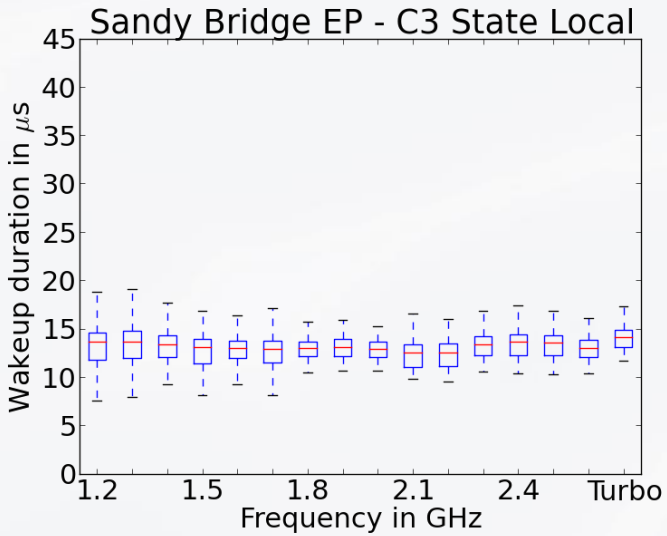
Results C3 (Intel, According to ACPI: 20/80 µs)

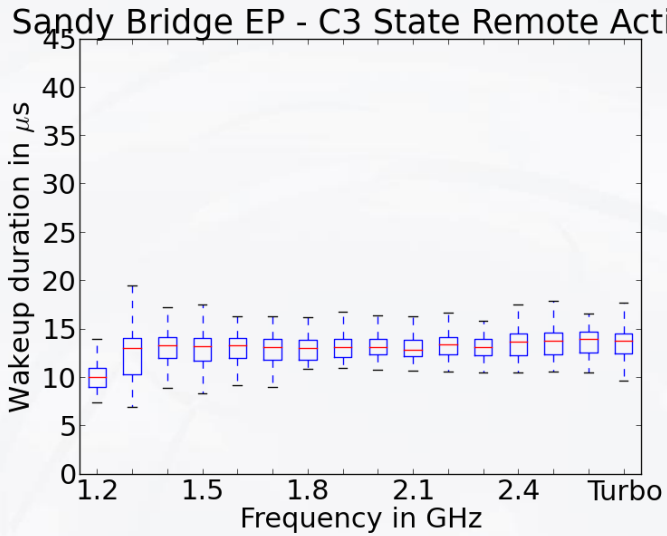
SPONSORED BY THE
Federal Ministry of Education and Research

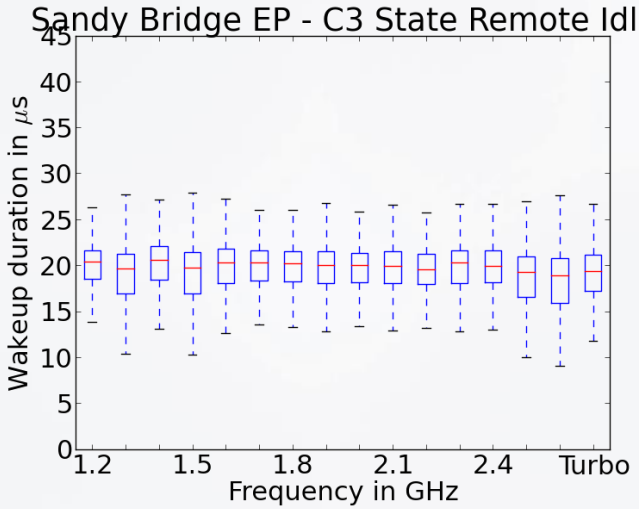


Results C3 (Intel, According to ACPI: 20/80 µs)

SPONSORED BY THE
Federal Ministry of Education and Research



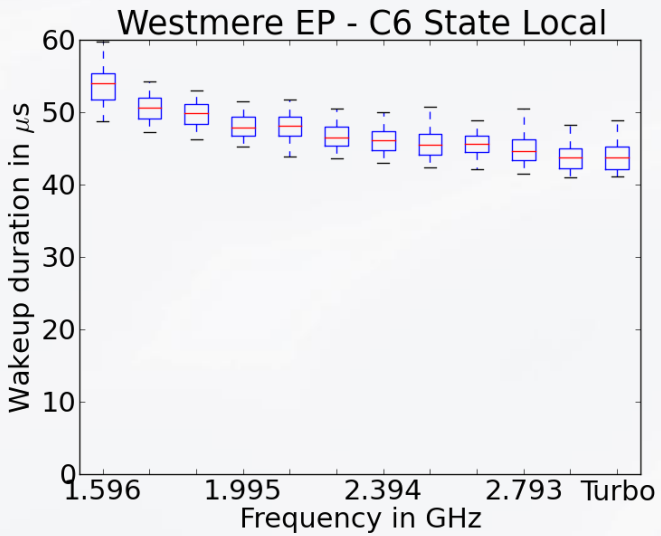




- Sandy Bridge ~20 µs faster than Westmere
- Package C3 adds approx. 6 µs in median
- Latency independent of frequency

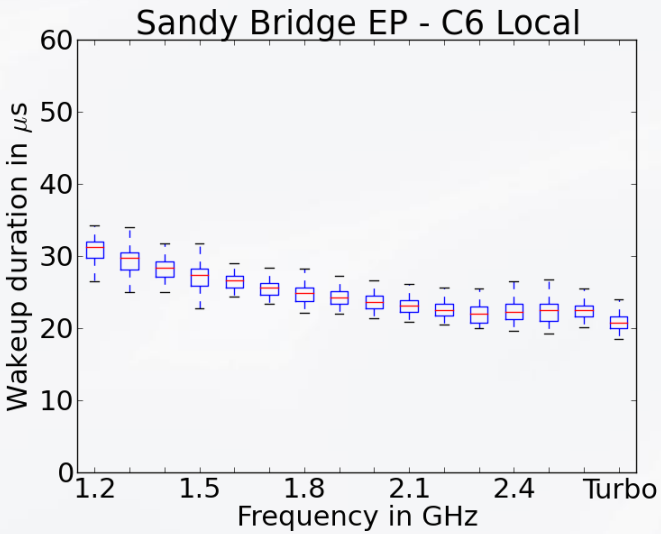
Results C6 (Intel, According to ACPI: 200/104 µs)

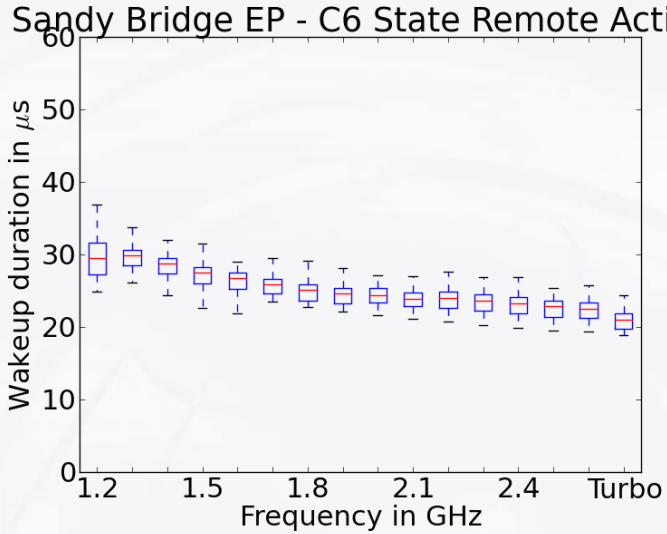
SPONSORED BY THE
Federal Ministry of Education and Research

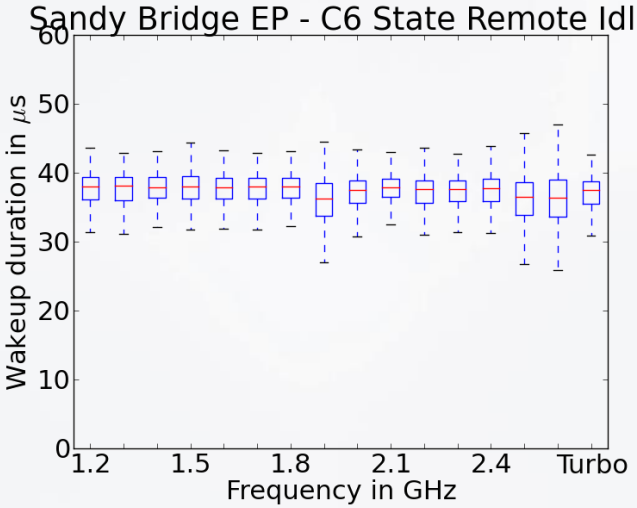


Results C6 (Intel, According to ACPI: 200/104 µs)

SPONSORED BY THE
Federal Ministry of Education and Research



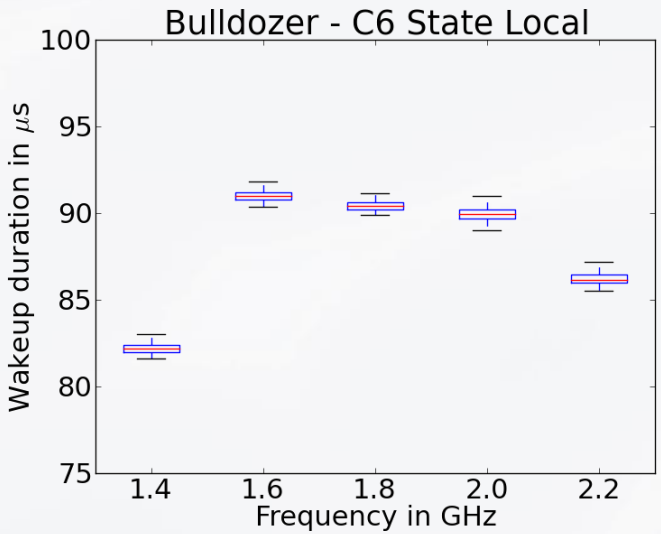


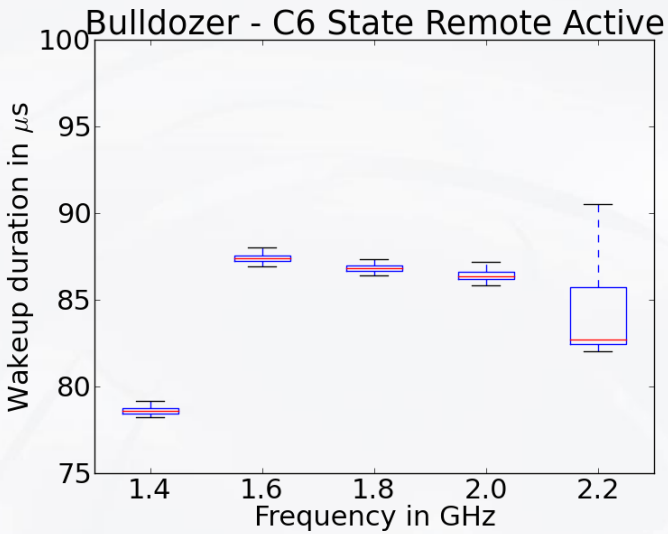


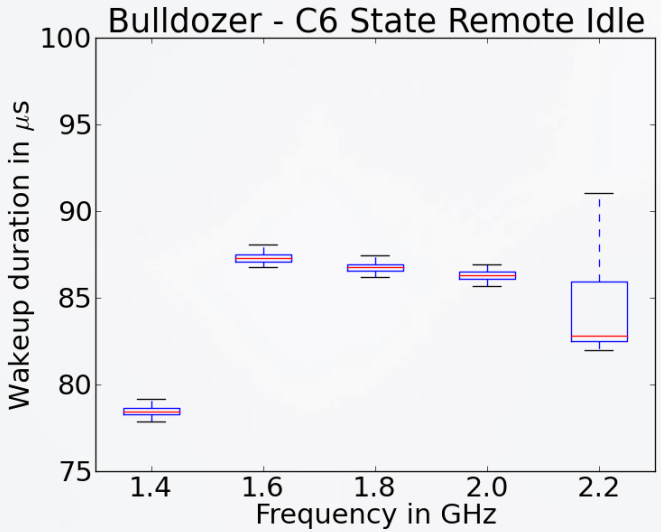
- Sandy Bridge ~20-13 µs faster than Westmere for C6
- C6 performance depends on frequency, PC6 does not

Results C6 (AMD, According to ACPI: 100 µs)

SPONSORED BY THE
Federal Ministry of Education and Research



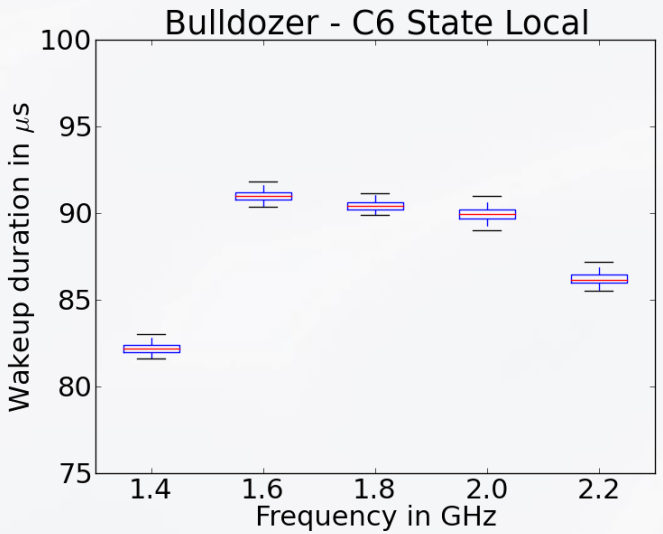


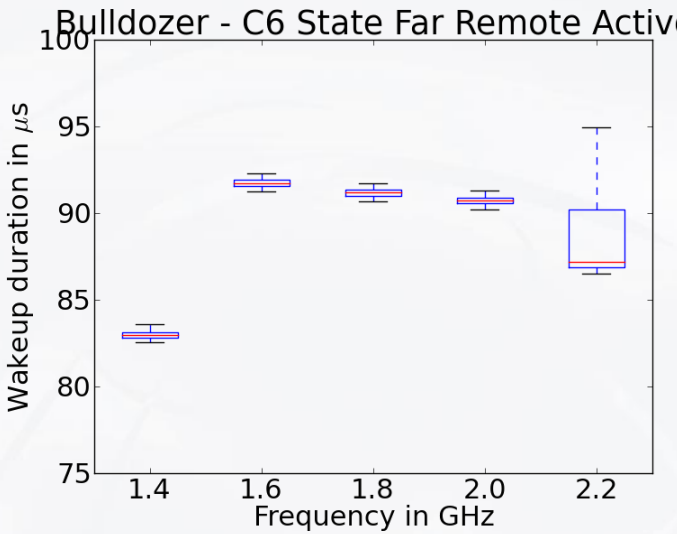


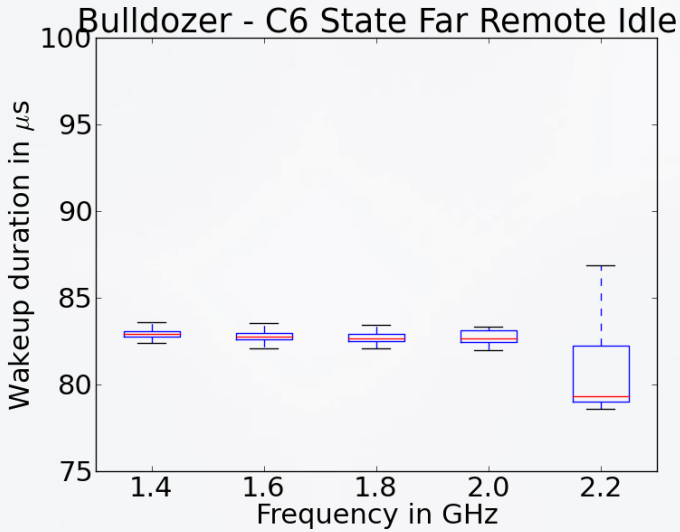
- Fastest on highest P-State
- Remote faster than local

Results C6 (AMD, According to ACPI: 100 µs)

SPONSORED BY THE
Federal Ministry of Education and Research







- Fastest on highest P-State
- Remote faster than local
- Only whole processors can do a voltage reduction, single dies cannot

Summary

SPONSORED BY THE
Federal Ministry of Education and Research

- ACPI projections too optimistic for Westmere and Bulldozer
- ACPI projections too pessimistic for Sandy Bridge
- OS uses wrong projections to choose best C-State

→ Redefine these values based on measurements and let OS know

- ACPI and OS unaware of dependencies between P- and C-States and Package C-States

Thank You

SPONSORED BY THE
Federal Ministry of Education and Research

Questions?

No word on power/energy saving?

SPONSORED BY THE
Federal Ministry of Education and Research

- Well this is something that depends!
  - On the processor frequency
  - On what you do, when you are in C0
    P(FIRESTARTER)>P(HPL)>P(while(1);)>P(sqrt(fp))
  - On what other devices contribute to the system power consumption
    - Idle(PC6)=75 W, Idle(PC3)=80 W Idle(C1E)=98 W Idle(C1)=137 W
    - Idle(PC6)=175 W, Idle(PC3)=180 W Idle(C1E)=198 W Idle(C1)=237 W
  - On how well your OS supports device power management
- Wrong impression if I would add such analysis for a specific system