HetCore: TFET-CMOS Hetero-Device Architecture for CPUs and GPUs

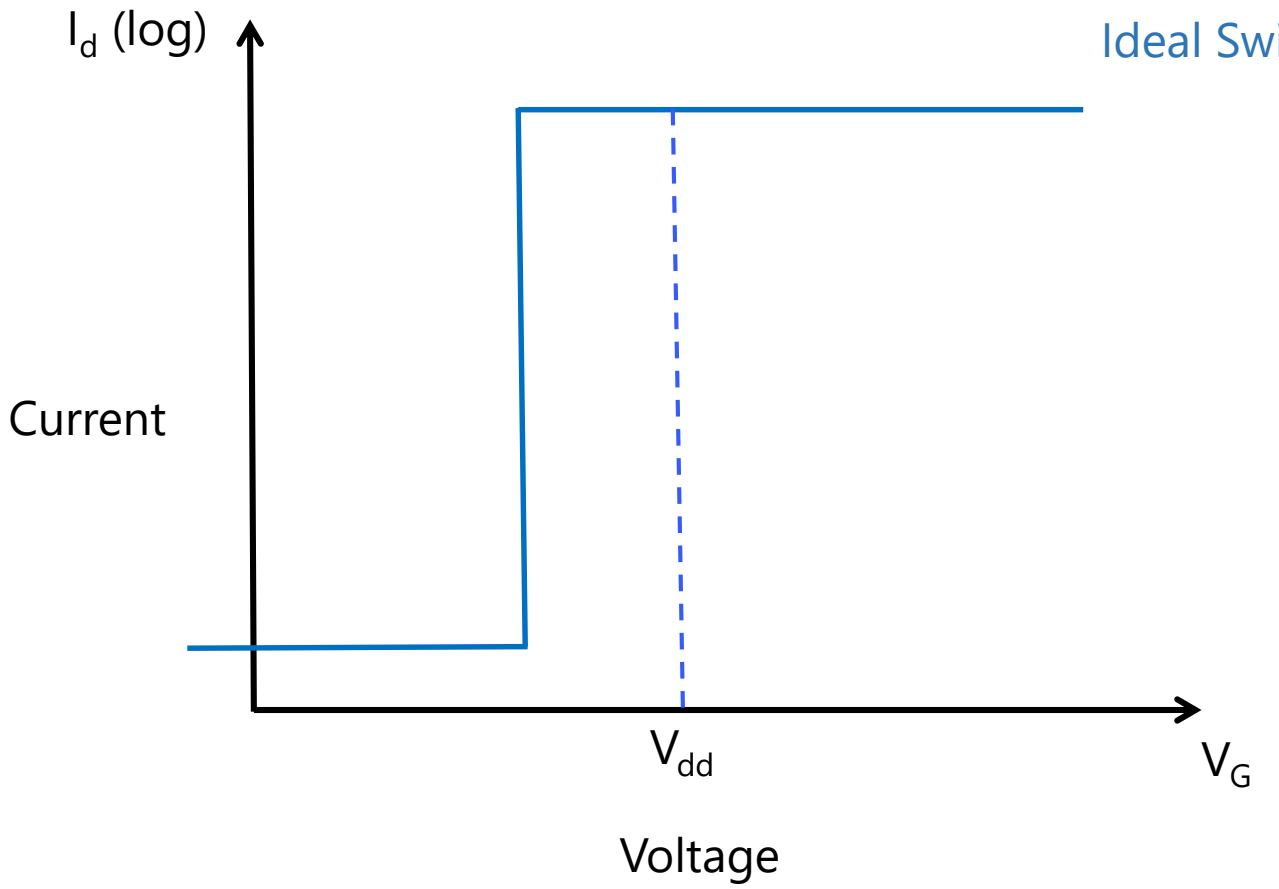
Bhargava Gopireddy, Dimitrios Skarlatos, Wenjuan Zhu, Josep Torrellas
University of Illinois at Urbana-Champaign
http://iacoma.cs.uiuc.edu

ISCA 2018
Wednesday, 11:20am
Session 9B: GPUs

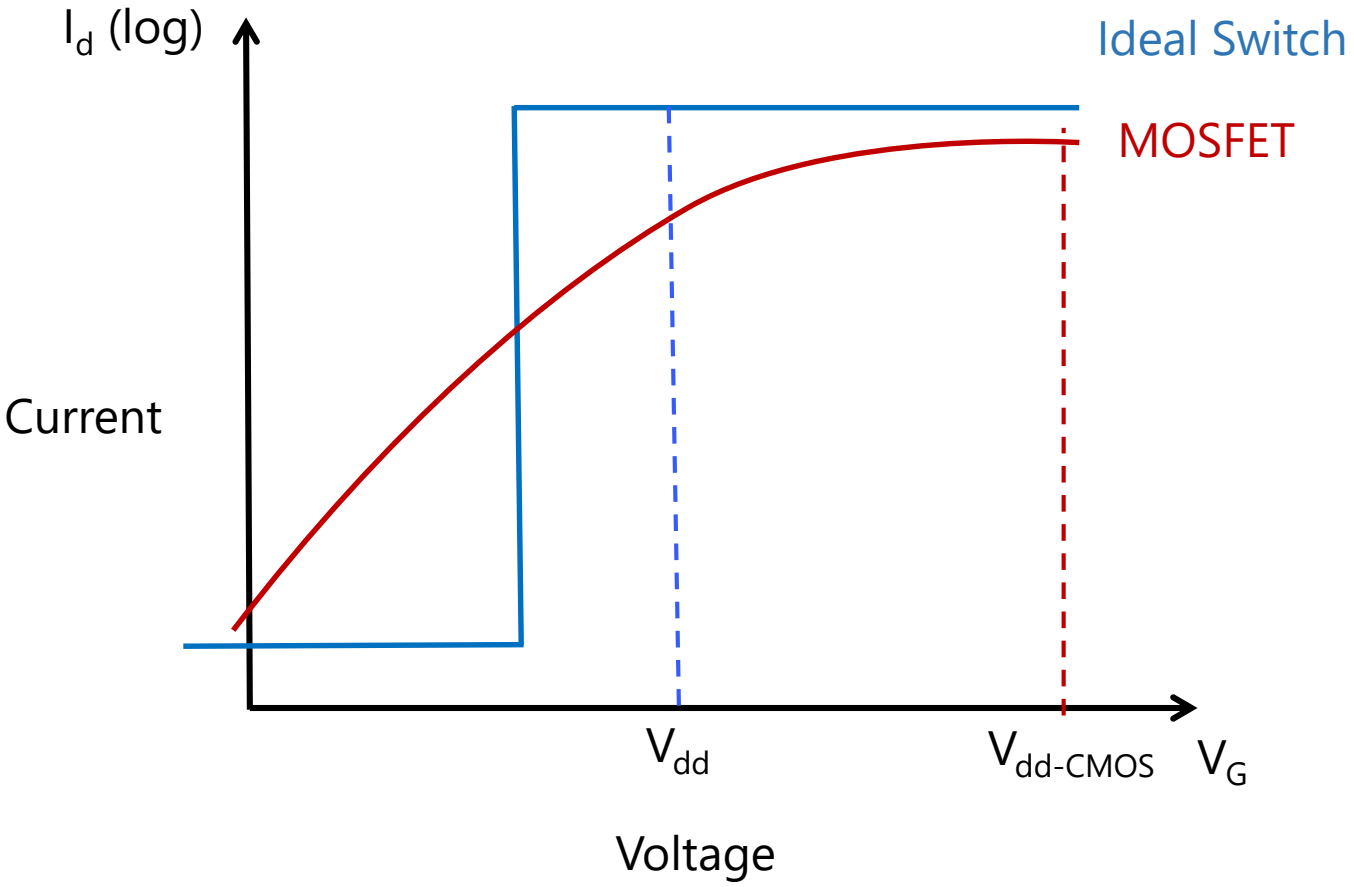
i-acoma group

ILLINOIS

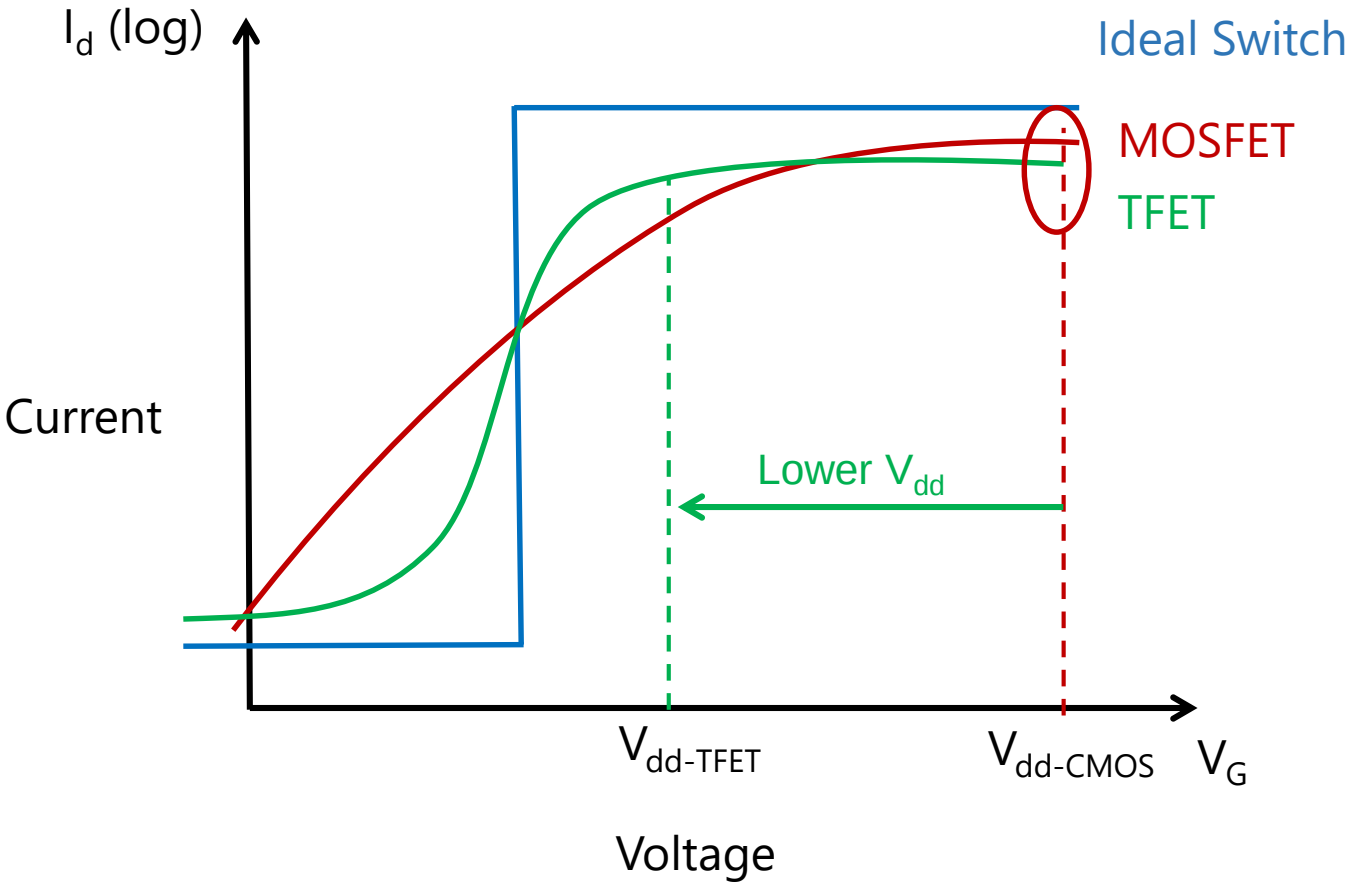
Ideal Switch



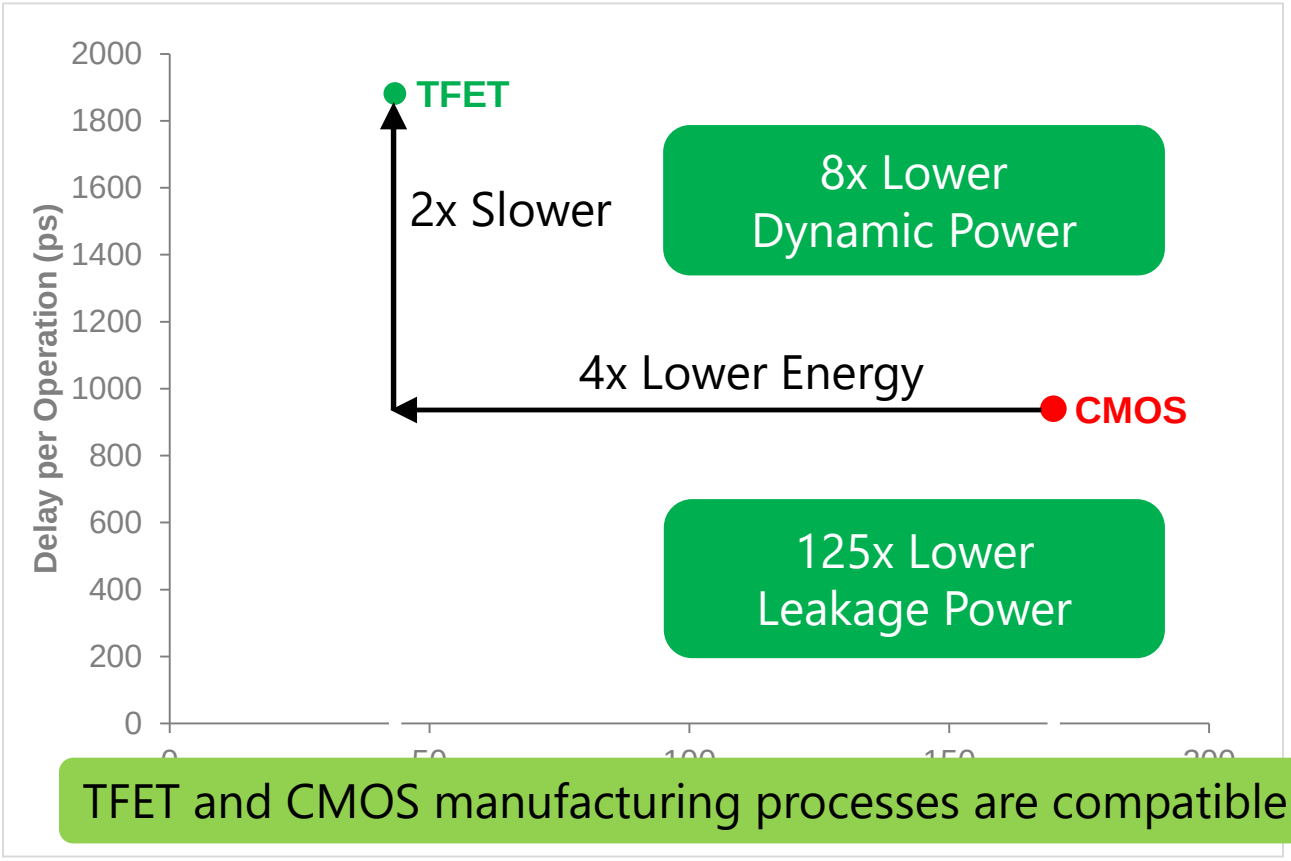
Ideal Switch vs Si-MOSFET



TFET vs MOSFET



TFET vs CMOS: Energy and Delay



V_{dd} at 15nm:
TFET: 0.4V
CMOS: 0.73V

TFET and CMOS manufacturing processes are compatible → Share same chip

Goal: Energy Efficient Core Design with TFETs

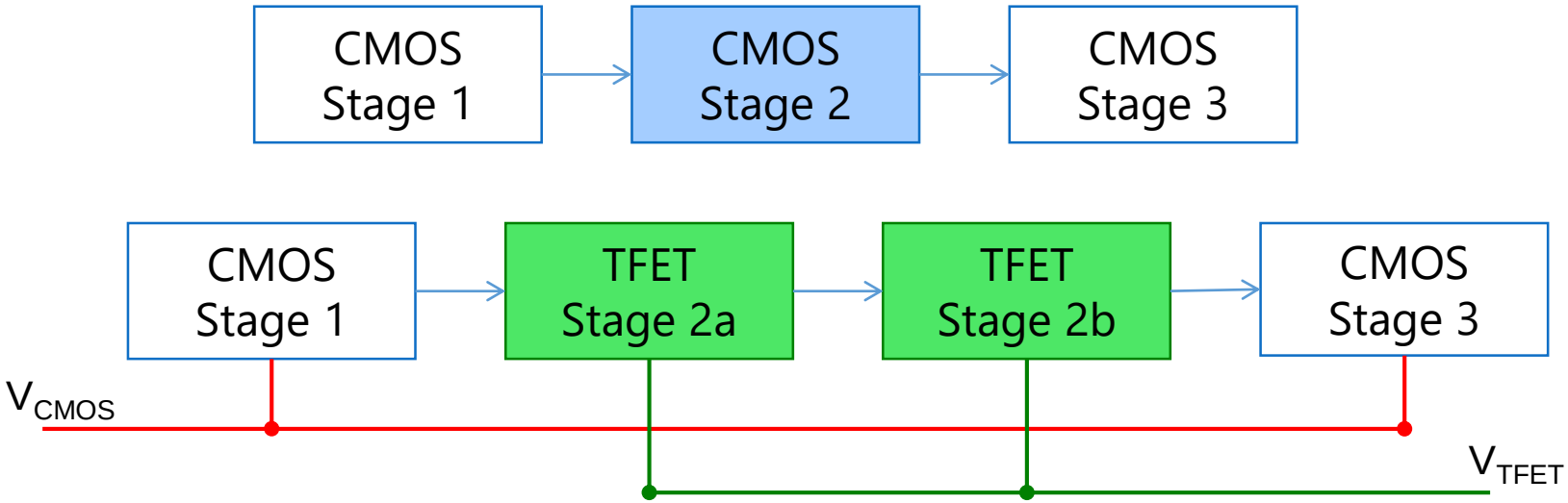
- Design a core that is
  - As energy efficient as TFET
  - As fast as CMOS
- Approach: Use both CMOS and TFET devices within the core
- How: Selectively replace CMOS units by TFET ones; that are
  - Power consuming
  - Amenable to pipelining or not very latency sensitive

Contributions

- Propose the concept of a hetero-device TFET-CMOS core architecture, called **HetCore**
- Design of an "Advanced HetCore" for CPUs and GPUs
  - Customizes known microarchitecture optimizations
- At iso-power, an 8-core HetCore CPU has a 68% lower ED^2 and is 32% faster than a 4-core CMOS CPU
- Similar results are obtained for GPUs

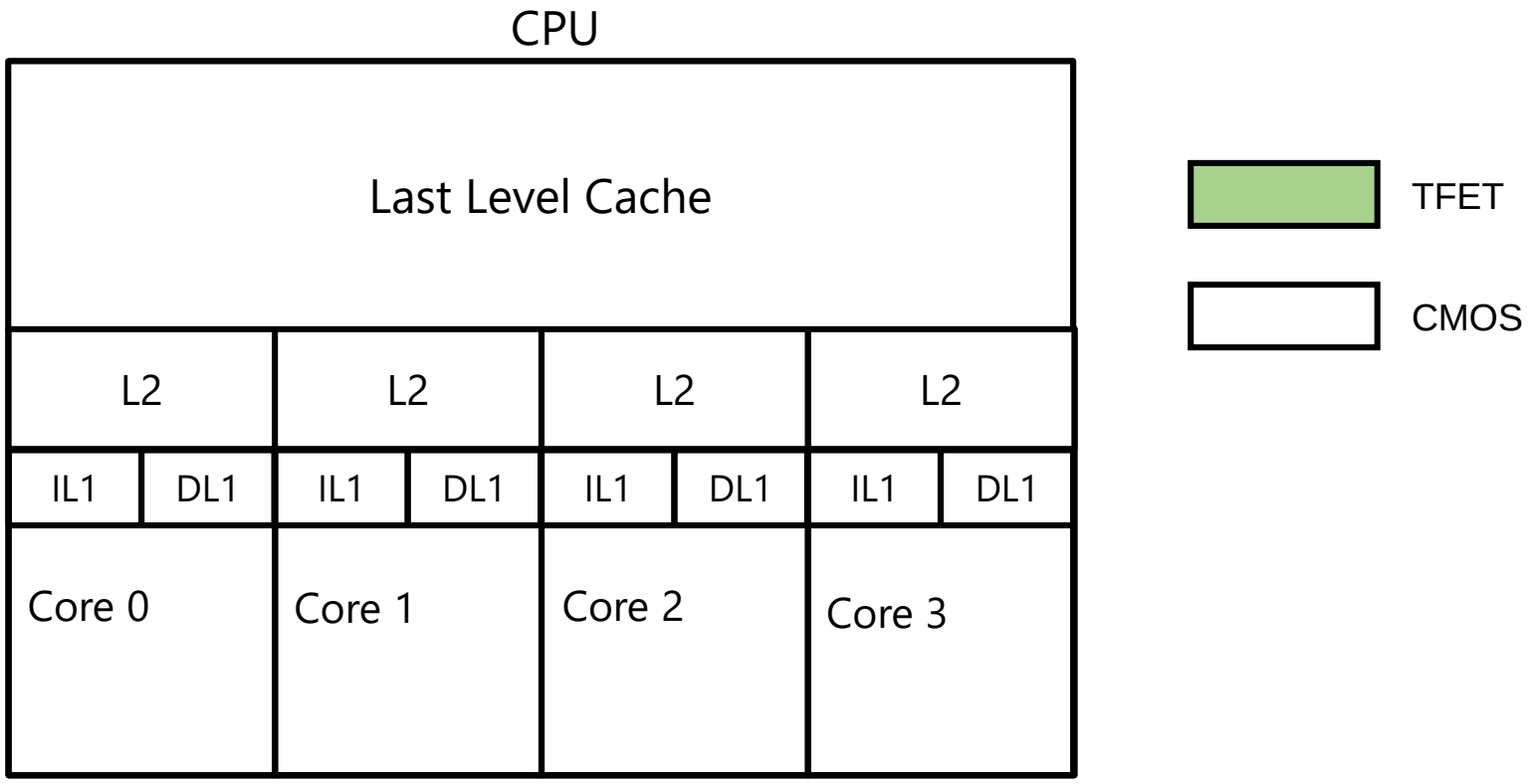
Replacing CMOS Units with TFET in Pipeline

- Pipeline twice as deep while maintaining the same frequency

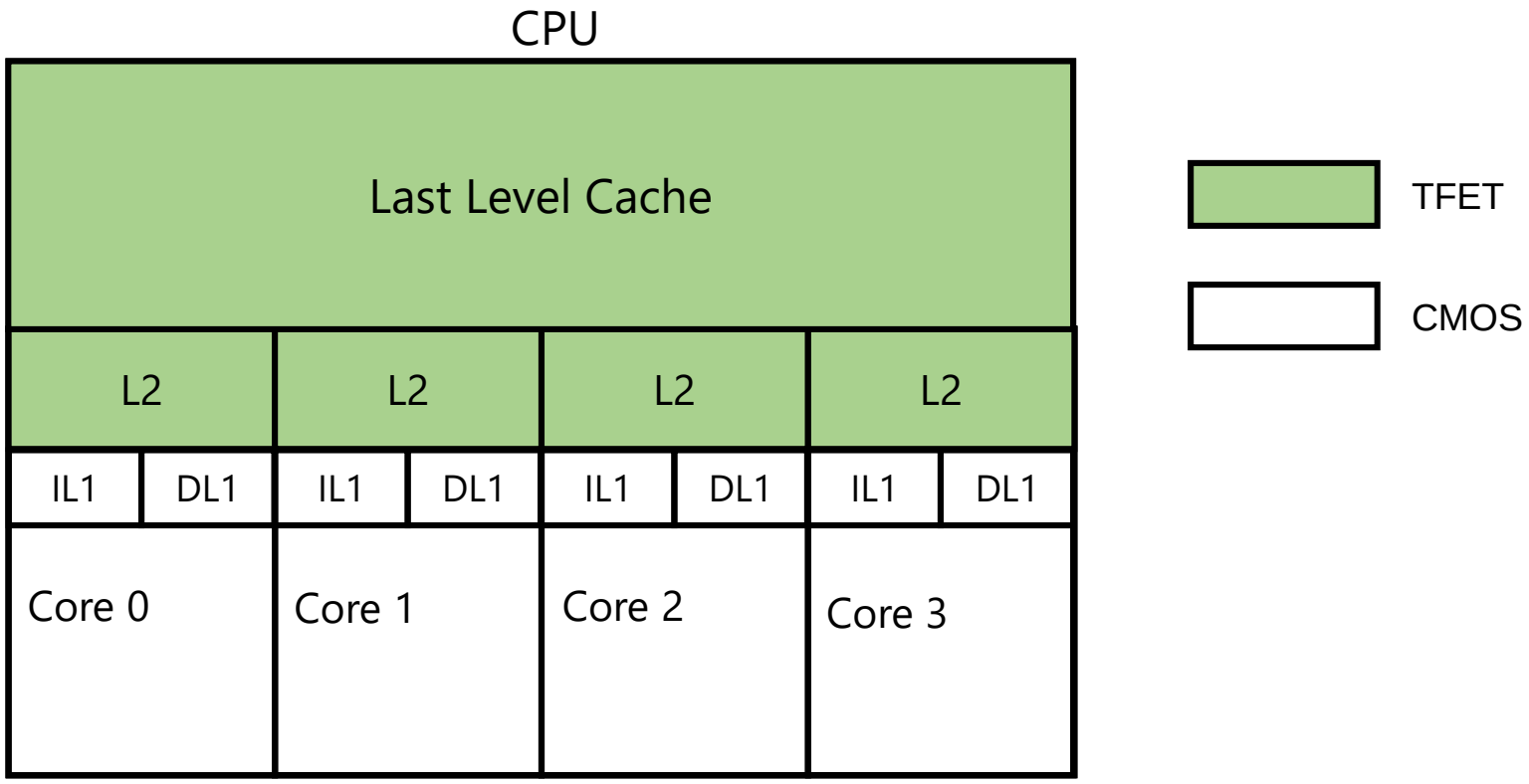


Selected units must be:
Amenable to pipelining and/or not very latency sensitive

Baseline HetCore Design

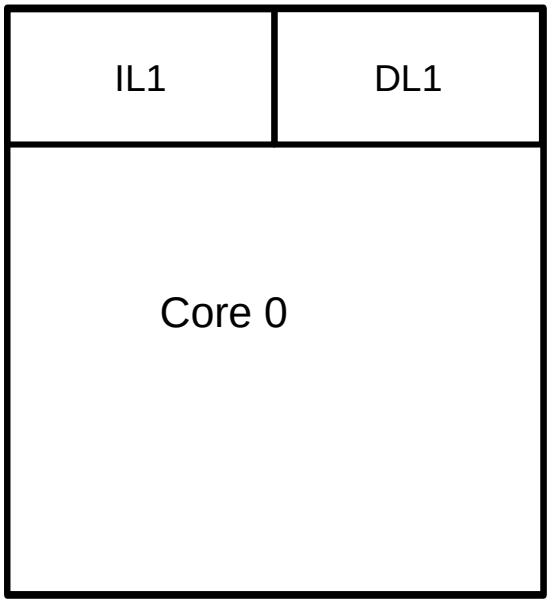


Baseline HetCore Design



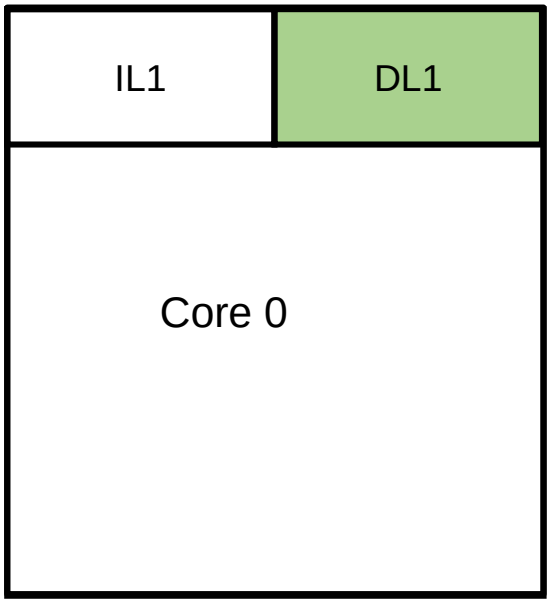
L2 and LLC primarily consume leakage power → TFETs can reduce leakage power substantially

Baseline HetCore Design



DL1 and IL1 consume high dynamic as well as leakage power

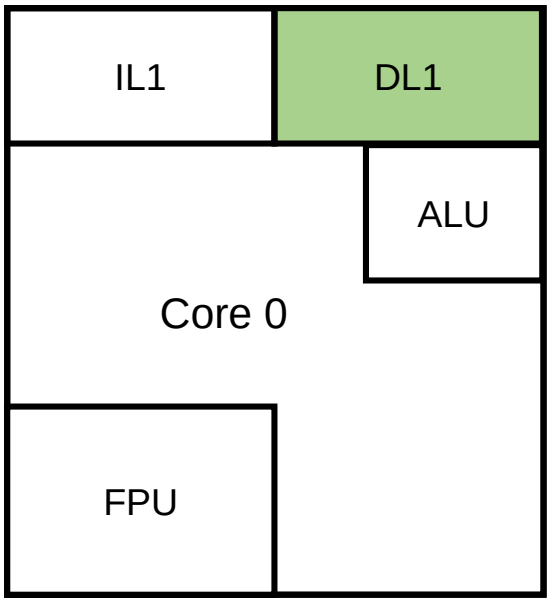
Baseline HetCore Design



DL1 and IL1 consume high dynamic as well as leakage power

DL1 latency can be partially hidden in an Out-of-Order machine

Baseline HetCore Design

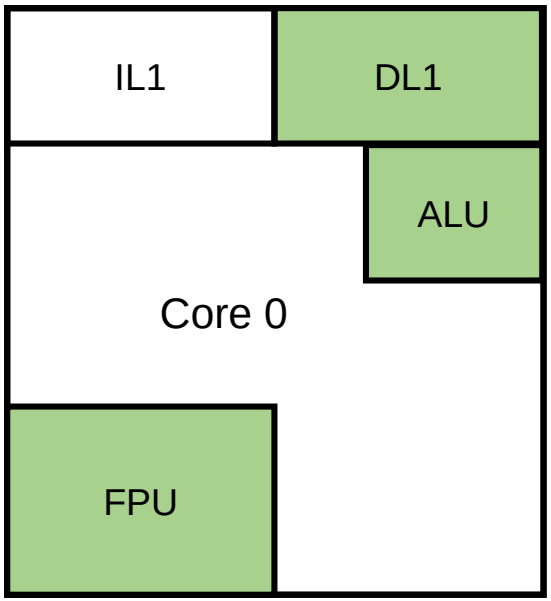


Both FPU and ALU consume significant power and can be pipelined

FPU: Pipeline deeper and exploit ILP

ALU: Impact on performance, but energy savings justify its placement in TFET

Baseline HetCore Design

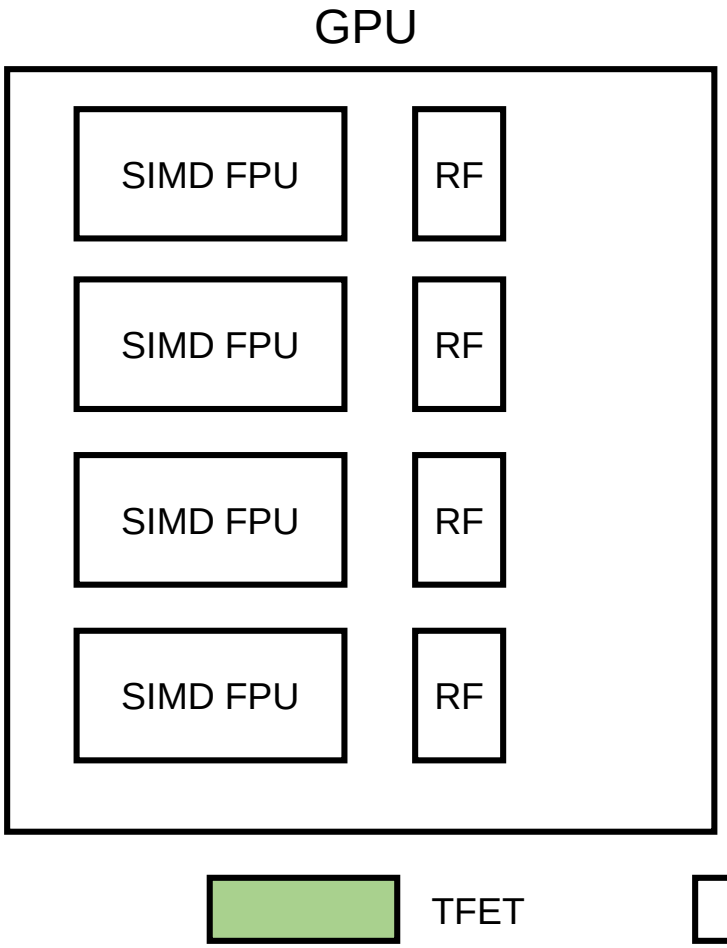


Both FPU and ALU consume significant power and can be pipelined

FPU: Pipeline deeper and exploit ILP

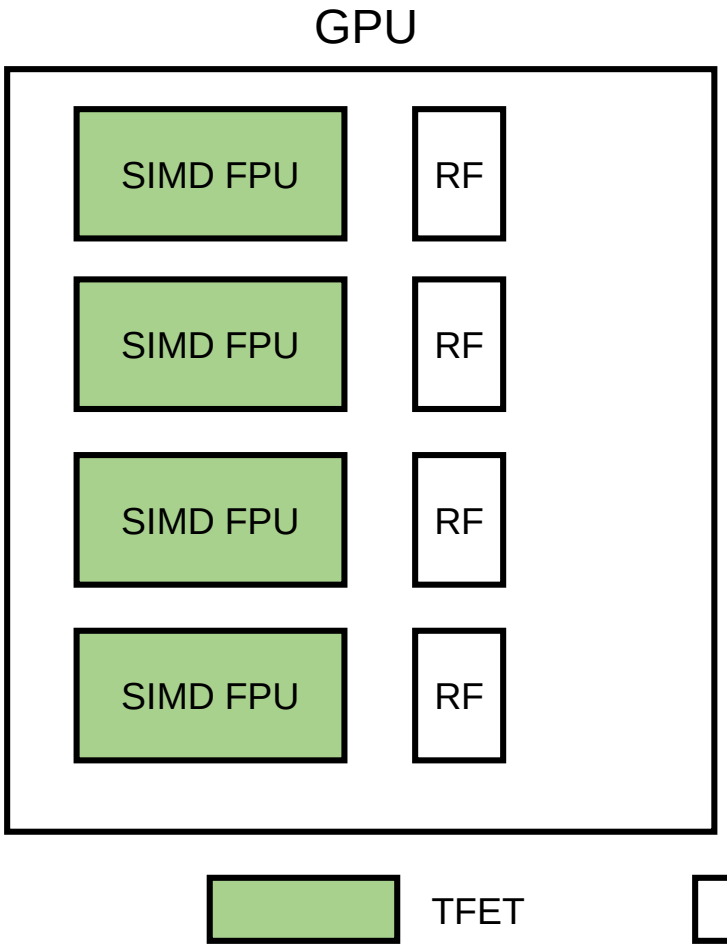
ALU: Impact on performance, but energy savings justify its placement in TFET

Baseline HetCore GPU Design



SIMD FPU can be pipelined

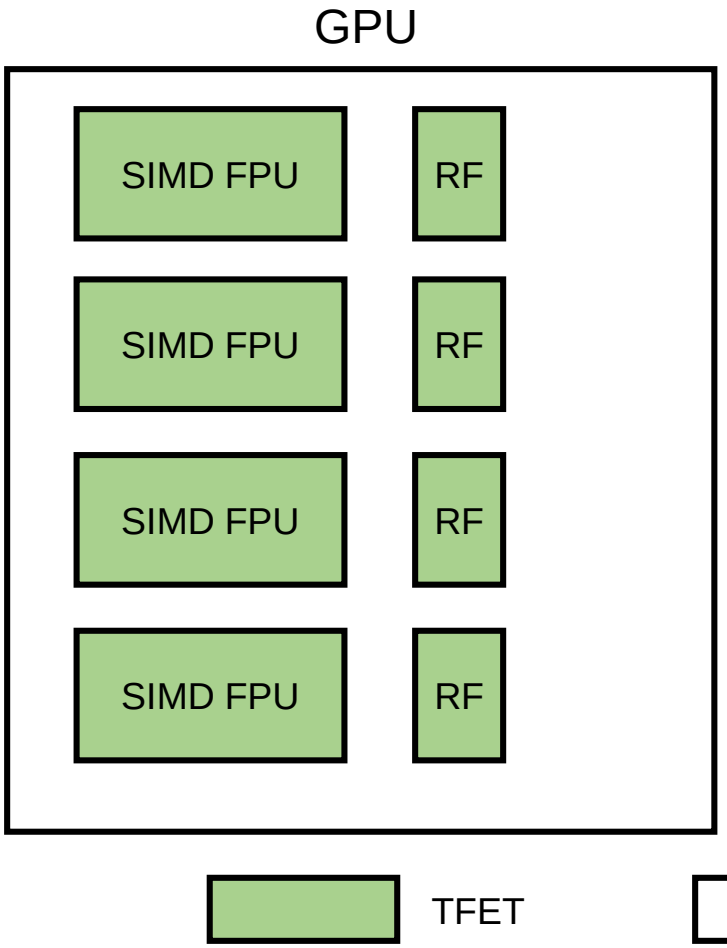
Baseline HetCore GPU Design



SIMD FPU can be pipelined

RF consumes high energy

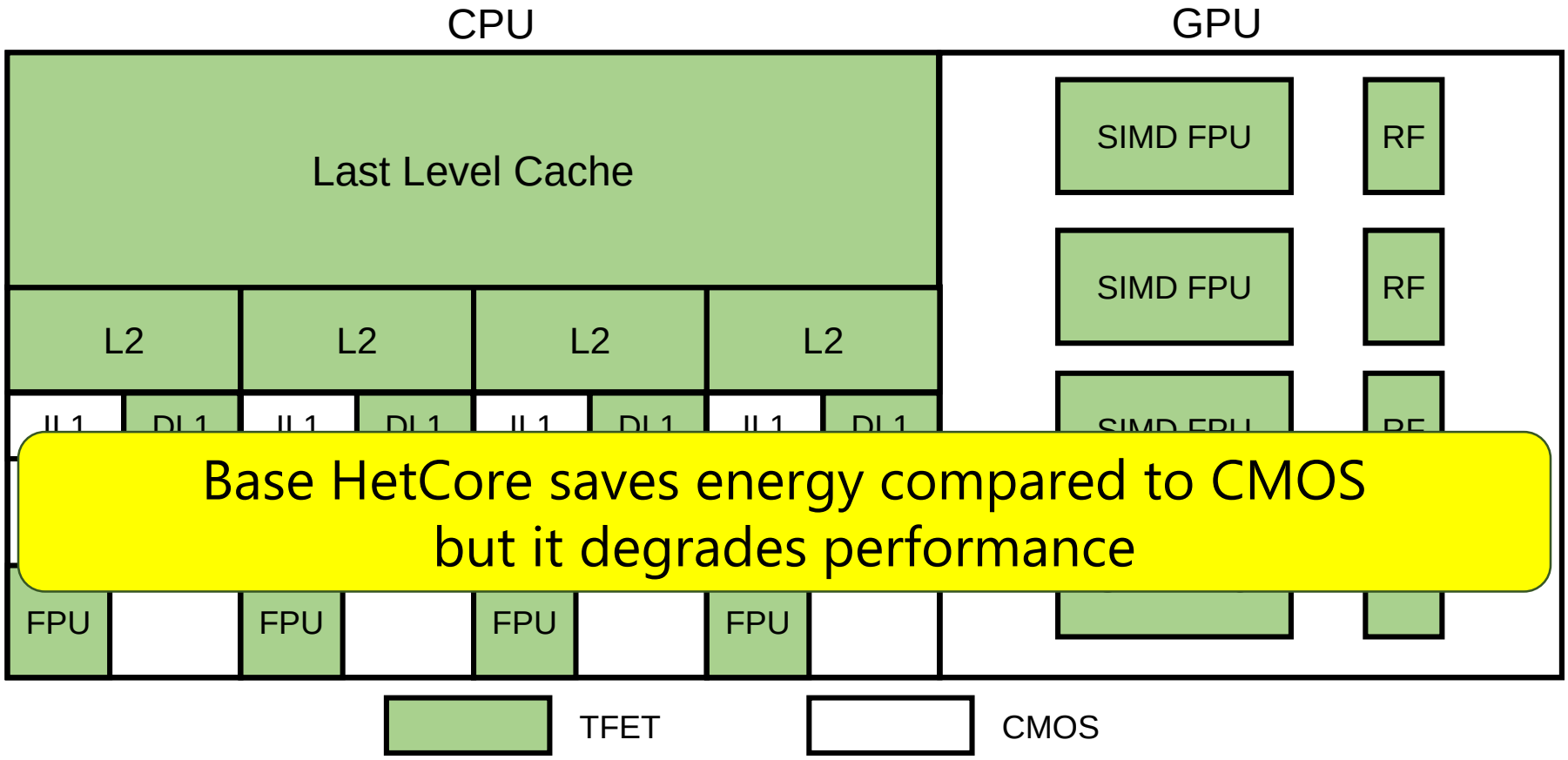
Baseline HetCore GPU Design



SIMD FPU can be pipelined

RF consumes high energy

Baseline HetCore with CPU and GPU



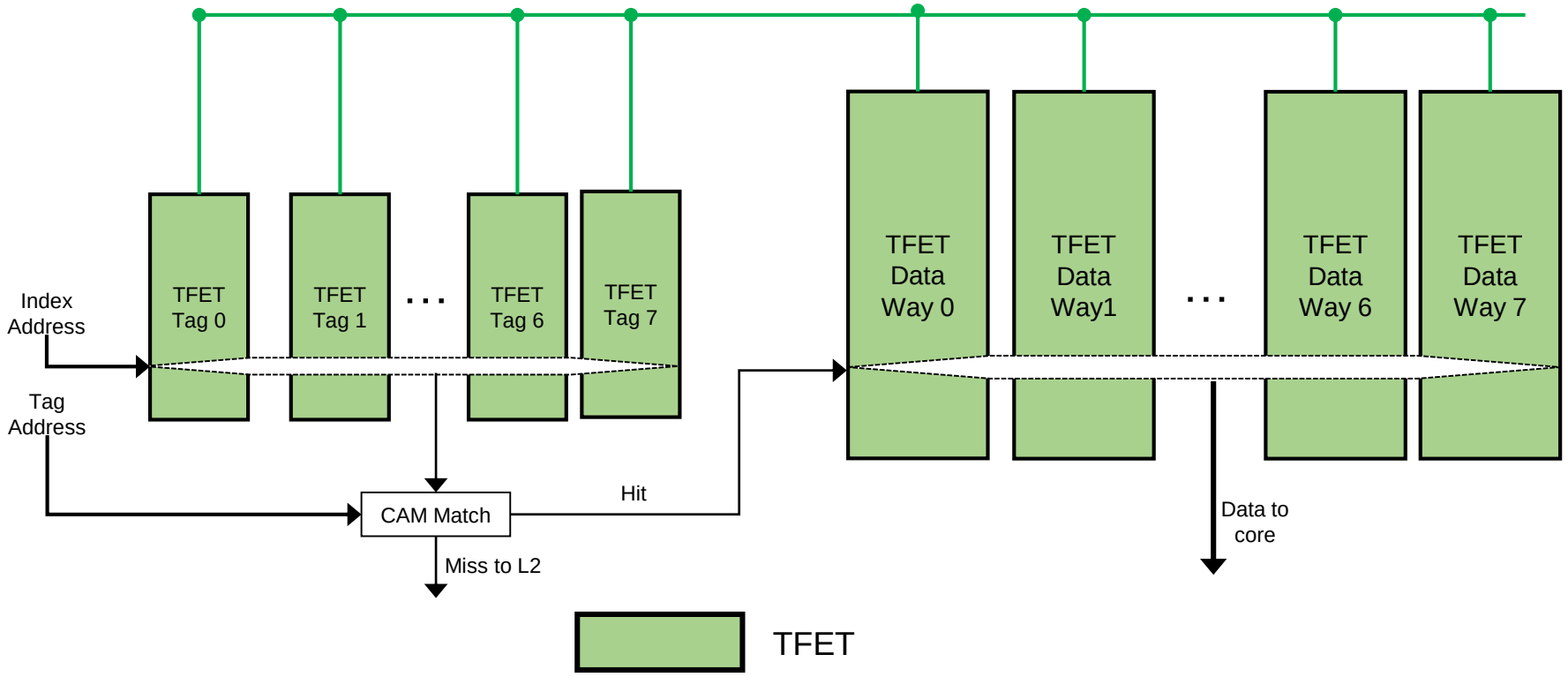
Base HetCore saves energy compared to CMOS but it degrades performance

Advanced HetCore Design

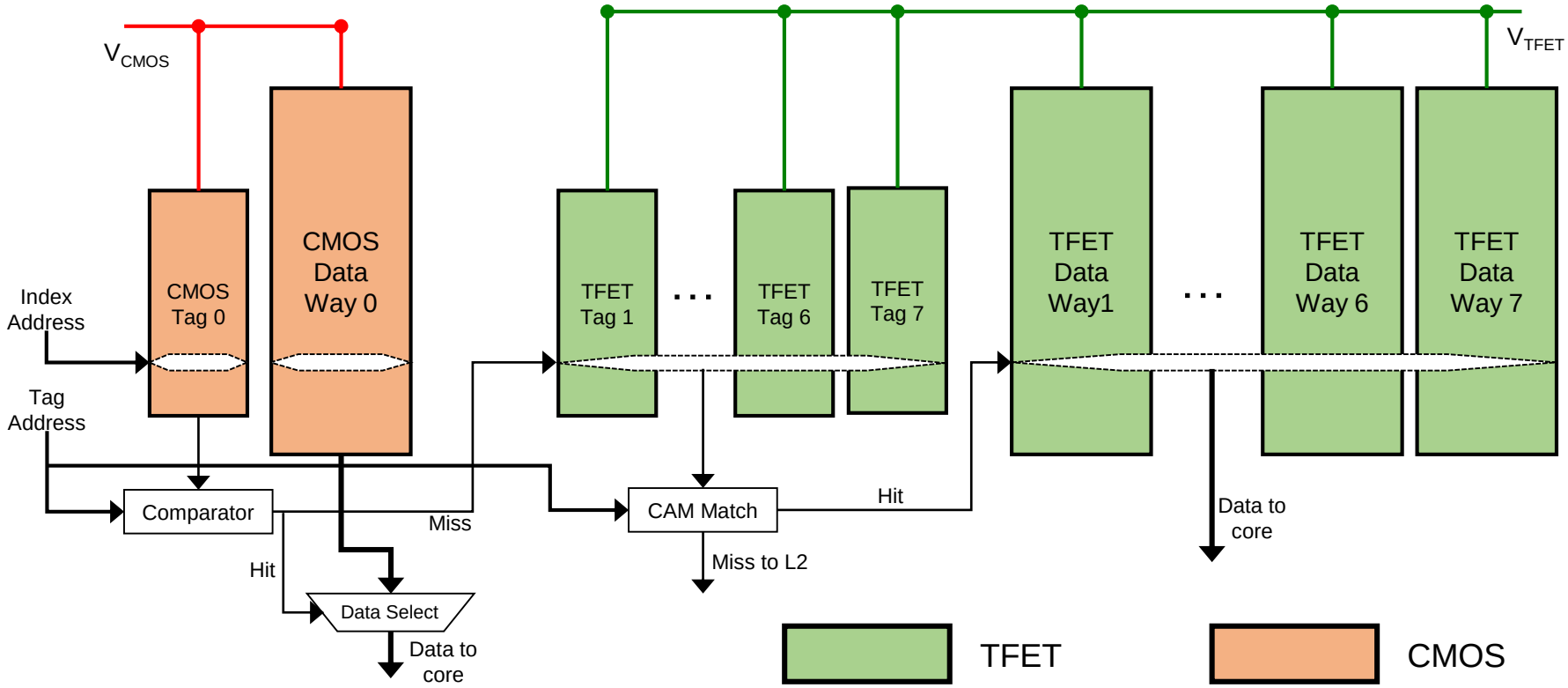
- New opportunities for micro-architectural optimization
  - Base HetCore is an unbalanced design
  - A small power penalty maybe a good tradeoff for large gains in performance

- For CPU:
  - Asymmetric DL1 cache
  - Dual cluster ALU
- For GPU:
  - Register file cache

DL1 Cache in TFET

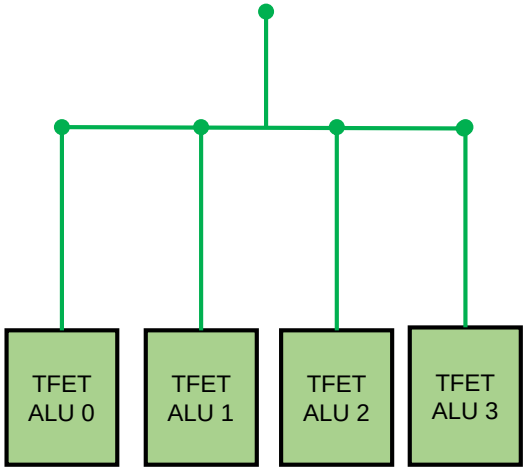


Asymmetric DL1 Cache



Check CMOS way before accessing TFET ways
CMOS way holds MRU cacheline and can respond in 1 cycle

Performance Impact of TFET ALU



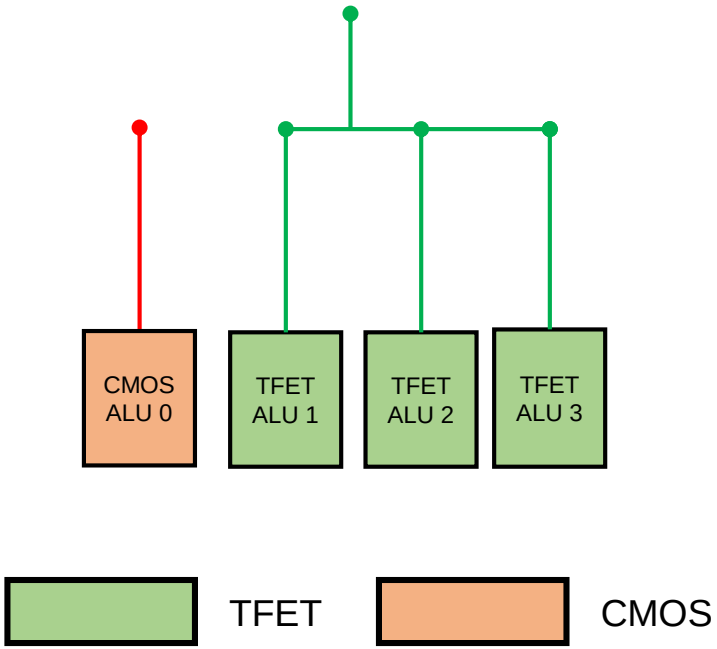
TFET

TFET ALU doubles the latency of most common operations

Prevents back-to-back issue of dependent instructions

Increases misprediction penalty

Dual Speed ALU Cluster



In dispatch stage, identify the producer-consumer pairs in small window, and steer the producer to CMOS ALU.

Steering algorithm: minimize bubbles, maximize power saving and balance overall utilization [Baniasadi et al]

Mis-steering a producer is okay; as the penalty is only one cycle for consumer

Register File Cache in GPU

- TFET register file introduces additional cycles in critical path
- Use: Register file cache, similar to an asymmetric cache, to hold a few registers closer to the FPU
  - Proposed earlier to reduce energy consumption [Gebhart et al.]
  - We use it to reduce the access latency by having the register file cache in CMOS

Evaluation Methodology

4 out-of-order cores in CPU, 8 Compute Units in GPU (AMD Southern Islands)

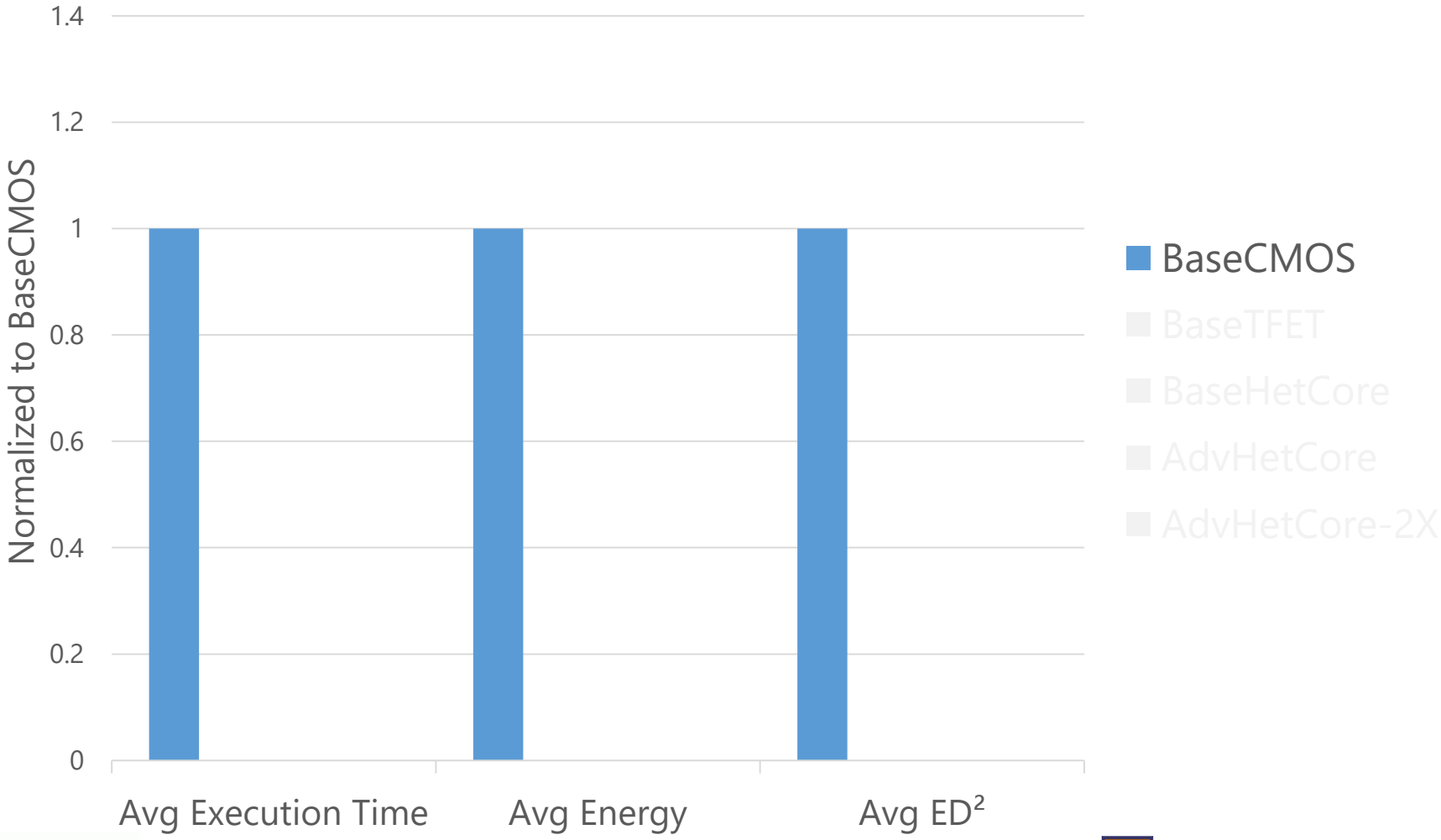
Multi2sim Simulator

- CPU: SPLASH2 and Parsec
- GPU: AMD-SDK-APP benchmark suite

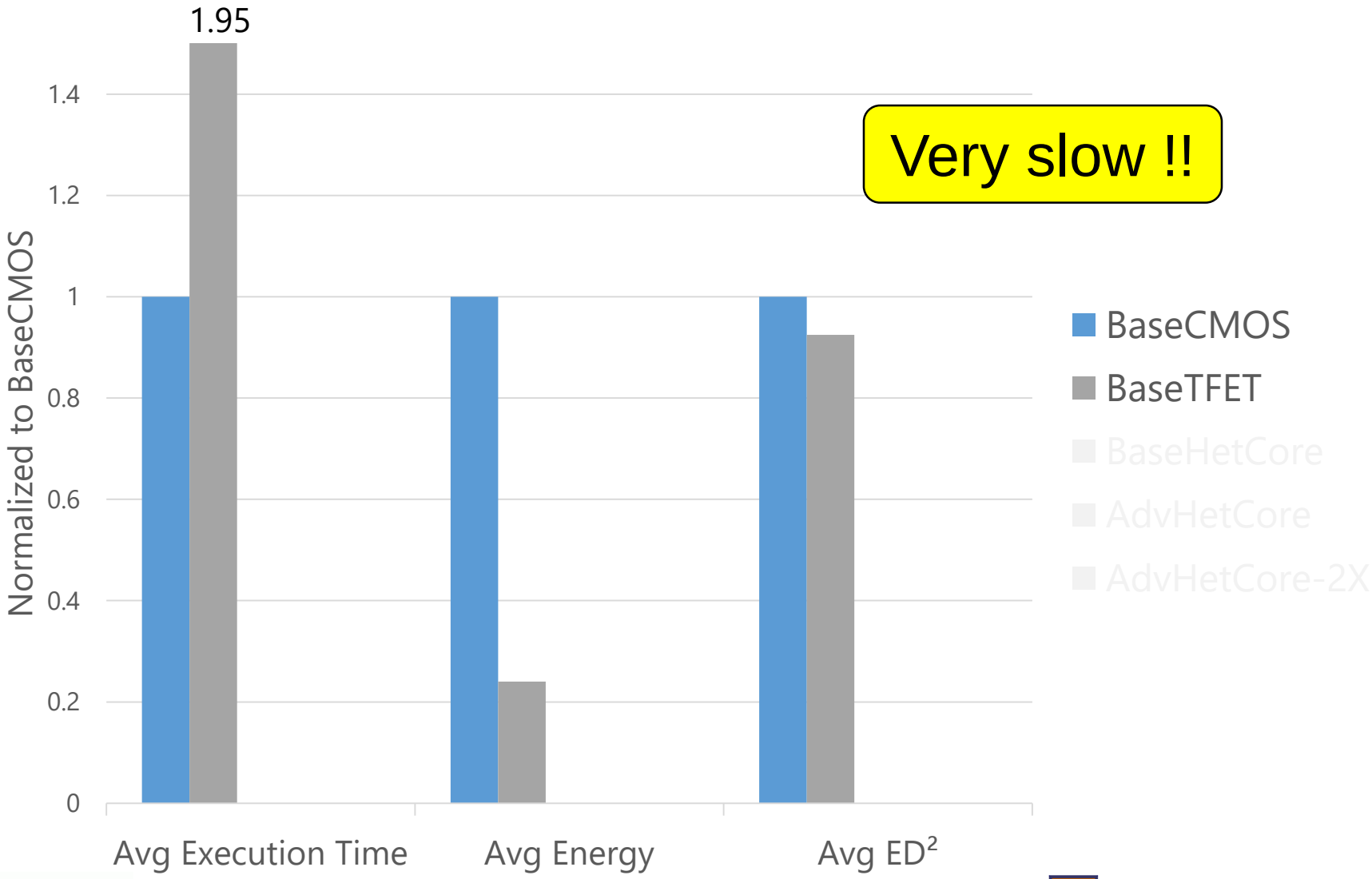
Configurations:

- BaseCMOS, BaseTFET
- Base HetCore
- Adv HetCore → Base HetCore with previous mitigations
- Adv HetCore-2X → Twice as many cores within the same power budget as BaseCMOS

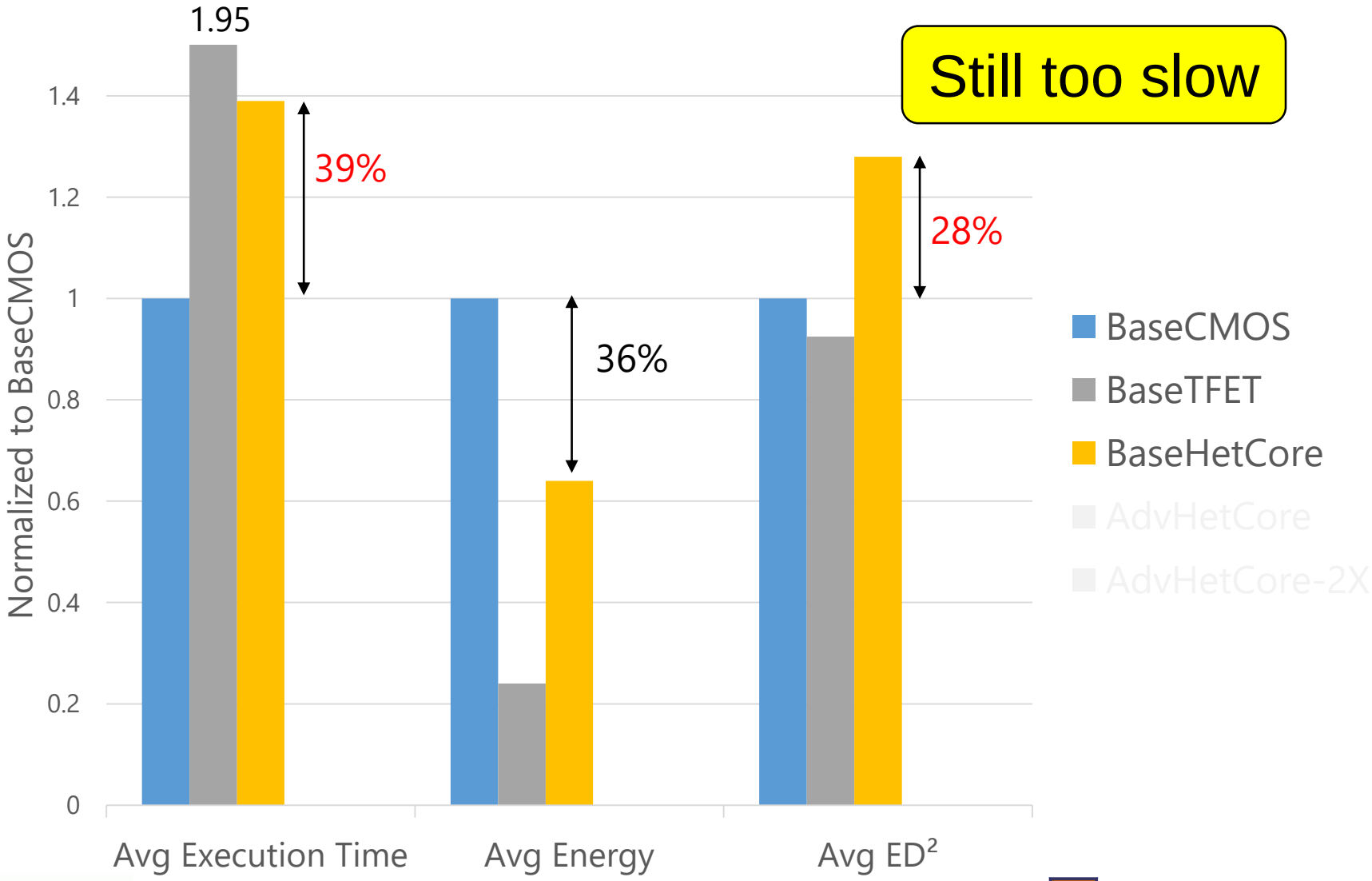
HetCore – CPU Results



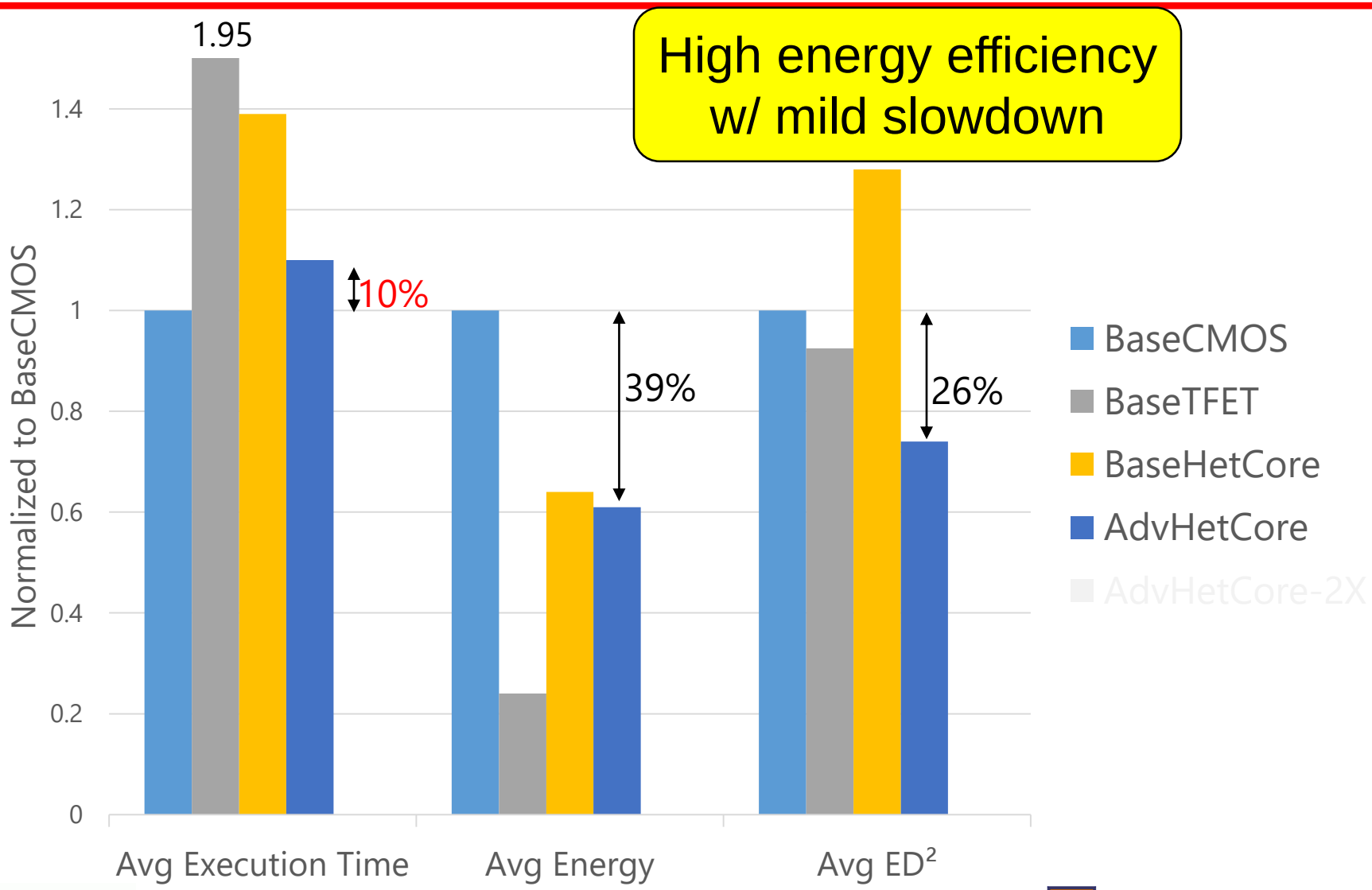
HetCore – CPU Results



Base HetCore – CPU Results

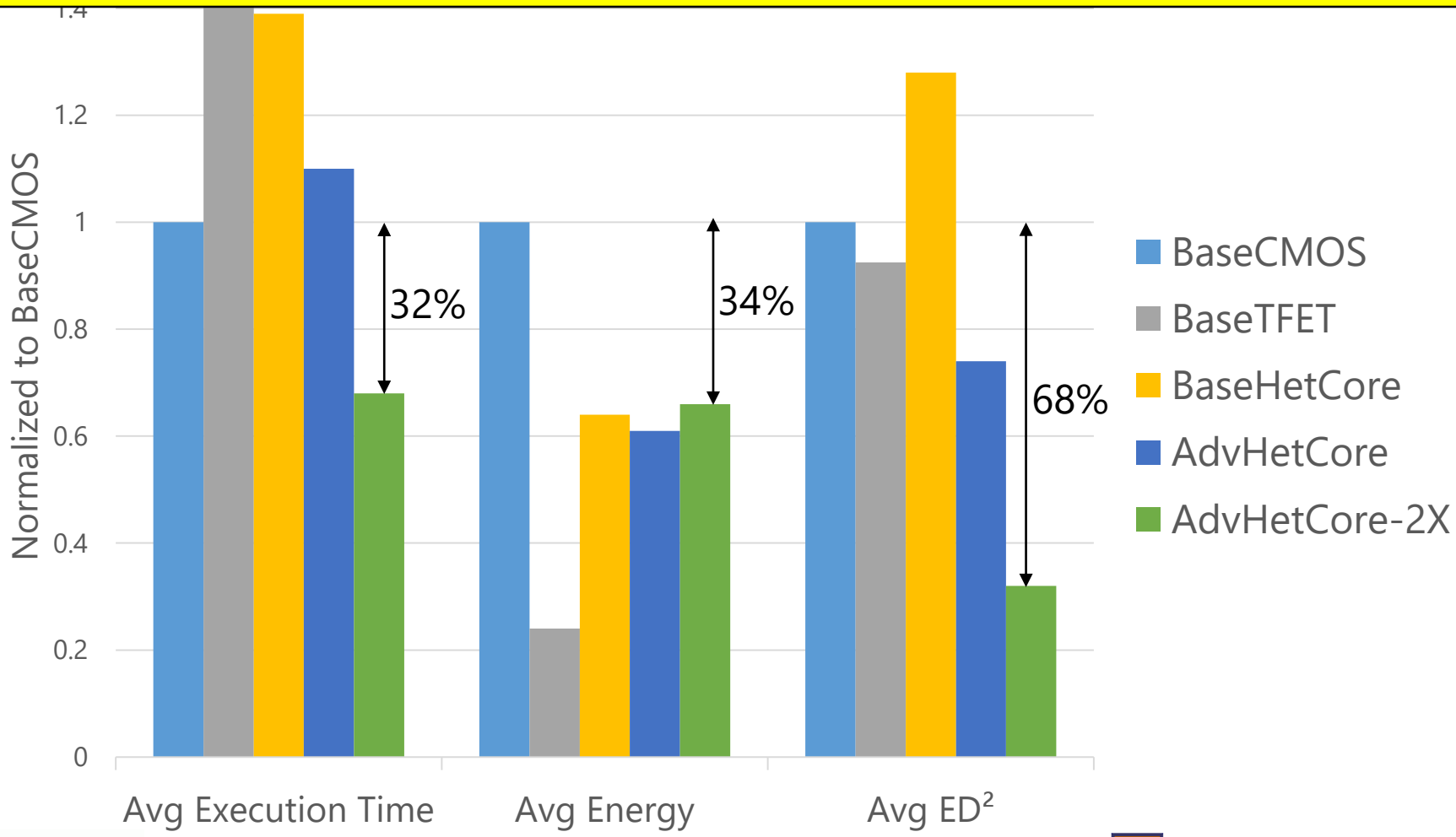


Adv HetCore – CPU Results



Adv HetCore-2X at Iso-power to BaseCMOS

Adv HetCore enables 2X cores in the same power budget !



Adv HetCore GPU

- Adv HetCore-GPU
  - 40% lower Energy
  - 20% slowdown

- Adv HetCore-GPU with 2X EUs at iso-power
  - 60% lower ED^2
  - 30% faster

Conclusion

- Proposed the concept of a hetero-device TFET-CMOS core architecture for high performance and energy efficiency
- Designed an Advanced HetCore for CPUs and GPUs
  - Customizes known microarchitecture optimizations
- At iso-power, an 8-core HetCore CPU has a 68% lower ED^2 and is 32% faster than a 4-core CMOS CPU
- Similar results are obtained for GPUs

HetCore: TFET-CMOS Hetero-Device Architecture for CPUs and GPUs

Bhargava Gopireddy, Dimitrios Skarlatos, Wenjuan Zhu, Josep Torrellas
University of Illinois at Urbana-Champaign
http://iacoma.cs.uiuc.edu

ISCA 2018
Wednesday, 11:20am
Session 9B: GPUs

i-acoma group

ILLINOIS