LeadOut: Composing Low-Overhead Techniques for Single-Thread Performance

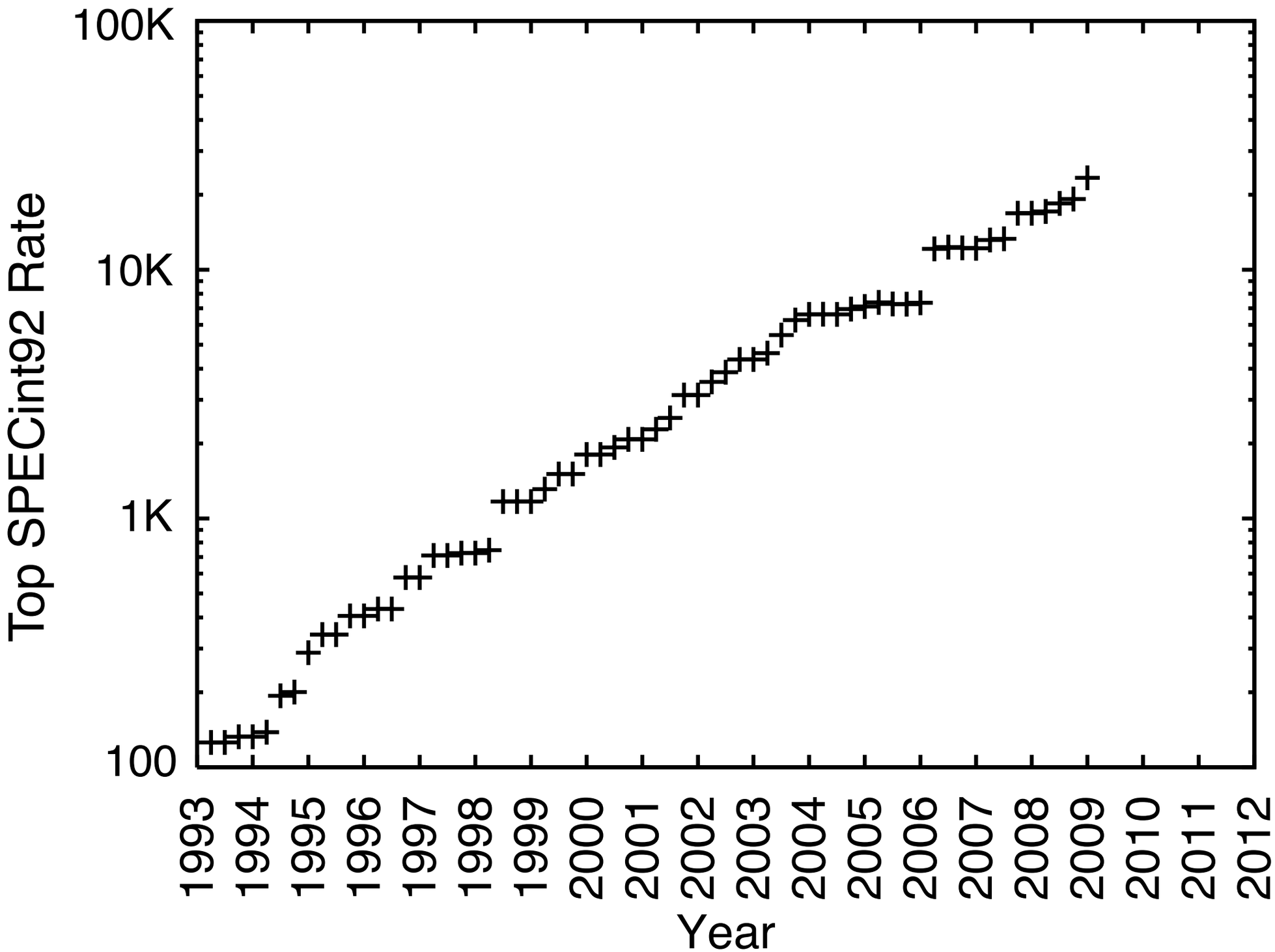
Brian Greskamp, **Ulya Karpuzcu**, Josep Torrellas

http://iacoma.cs.uiuc.edu/

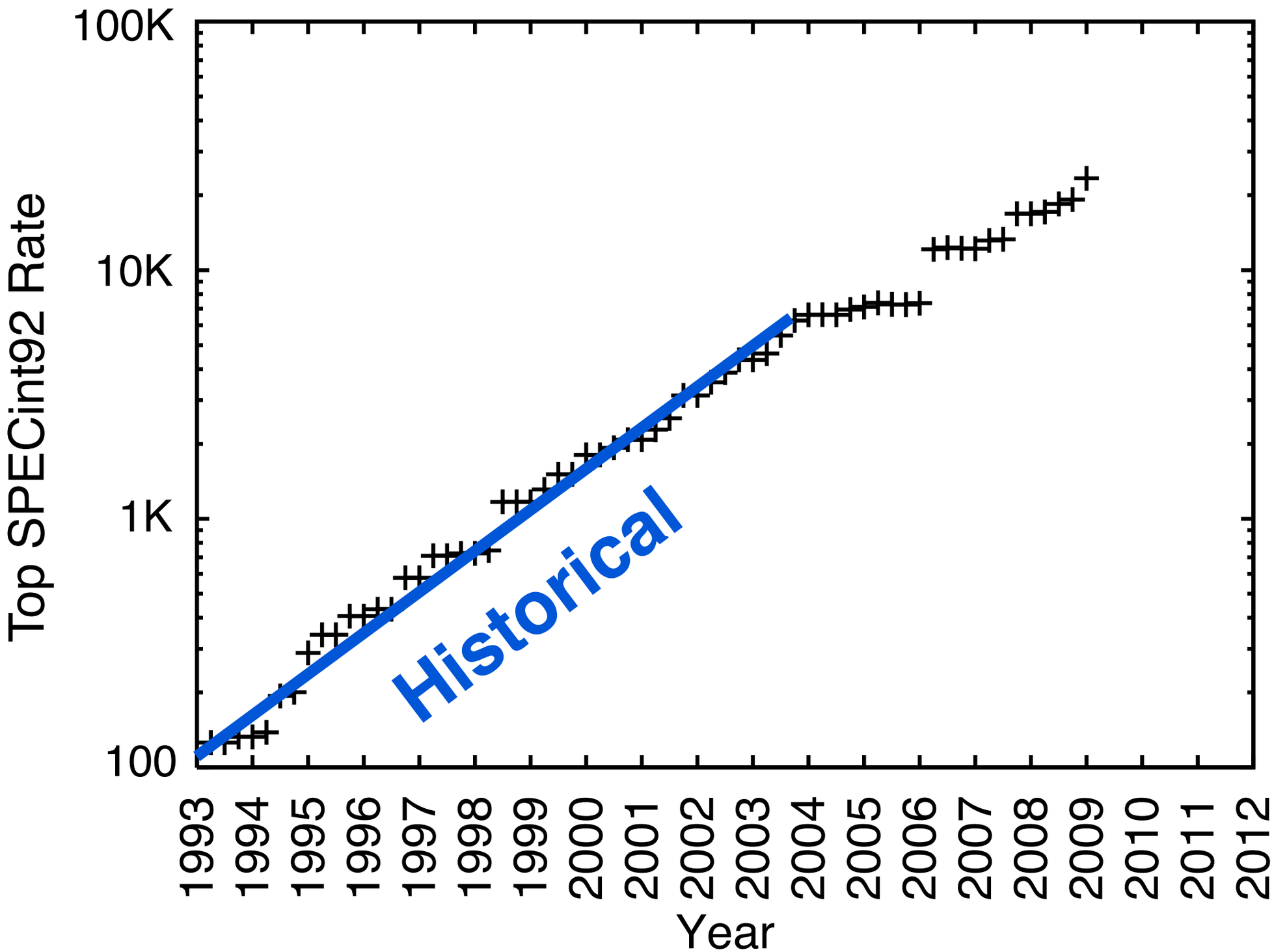
University
of Illinois

i-acoma group

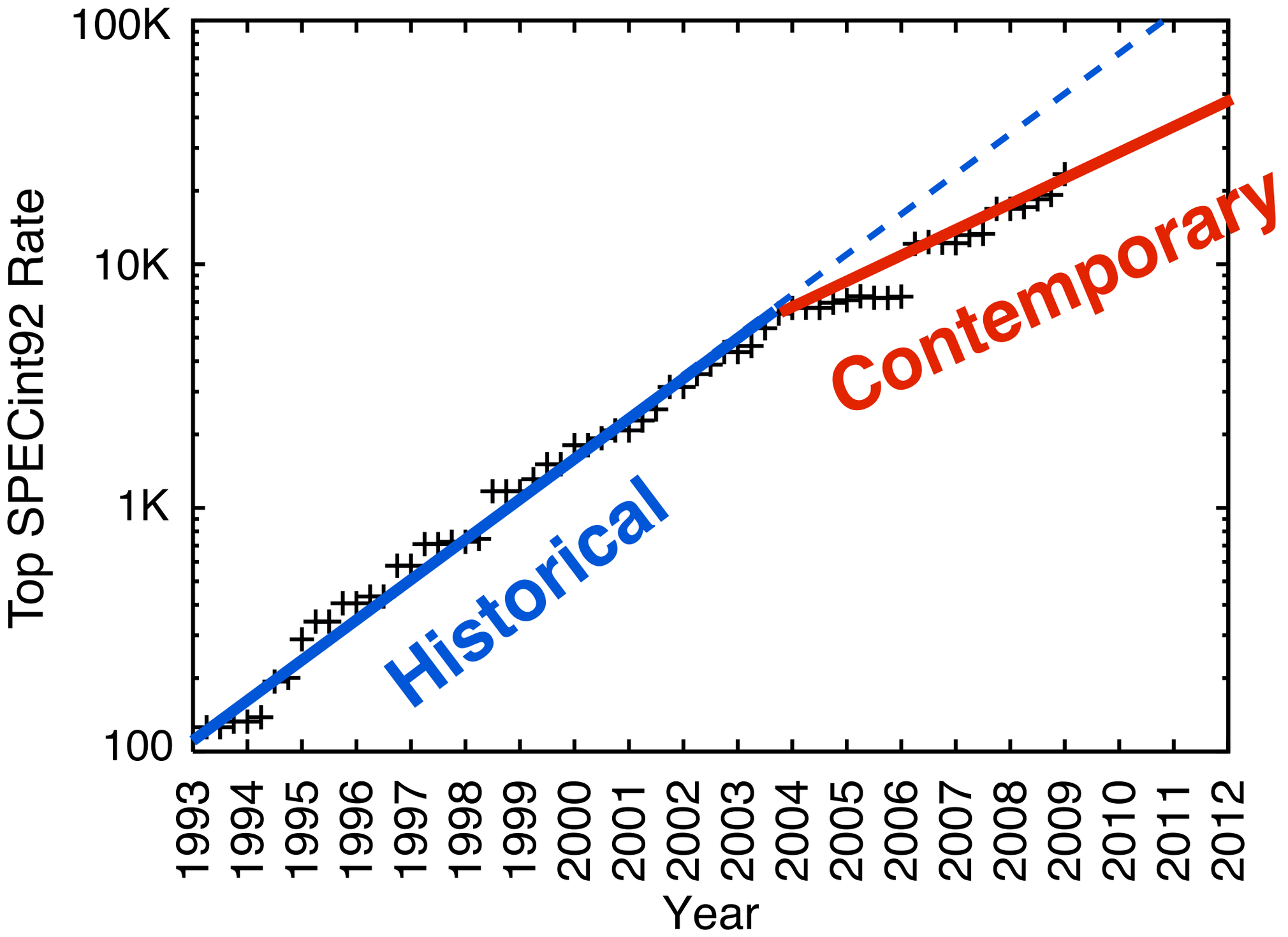
Per-Thread Performance Trend



Per-Thread Performance Trend



Per-Thread Performance Trend



Near-Future CMP Environment

Near-Future CMP Environment

- Sequential applications need **fast** cores

Near-Future CMP Environment

- Sequential applications need **fast** cores
- Throughput applications demand **more** cores

Near-Future CMP Environment

- Sequential applications need **fast** cores
- Throughput applications demand **more** cores
- Amdahl's Law: Most applications need some fast cores

Near-Future CMP Environment

- Sequential applications need **fast** cores
- Throughput applications demand **more** cores
- Amdahl's Law: Most applications need some fast cores
- Faster cores without compromising core count?

Near-Future CMP Environment

- Sequential applications need **fast** cores
- Throughput applications demand **more** cores
- Amdahl's Law: Most applications need some fast cores
- Faster cores without compromising core count?
  - Configurable Timing Speculation

Near-Future CMP Environment

- Sequential applications need **fast** cores
- Throughput applications demand **more** cores
- Amdahl's Law: Most applications need some fast cores
- Faster cores without compromising core count?
  - Configurable Timing Speculation
  - V/f Boosting

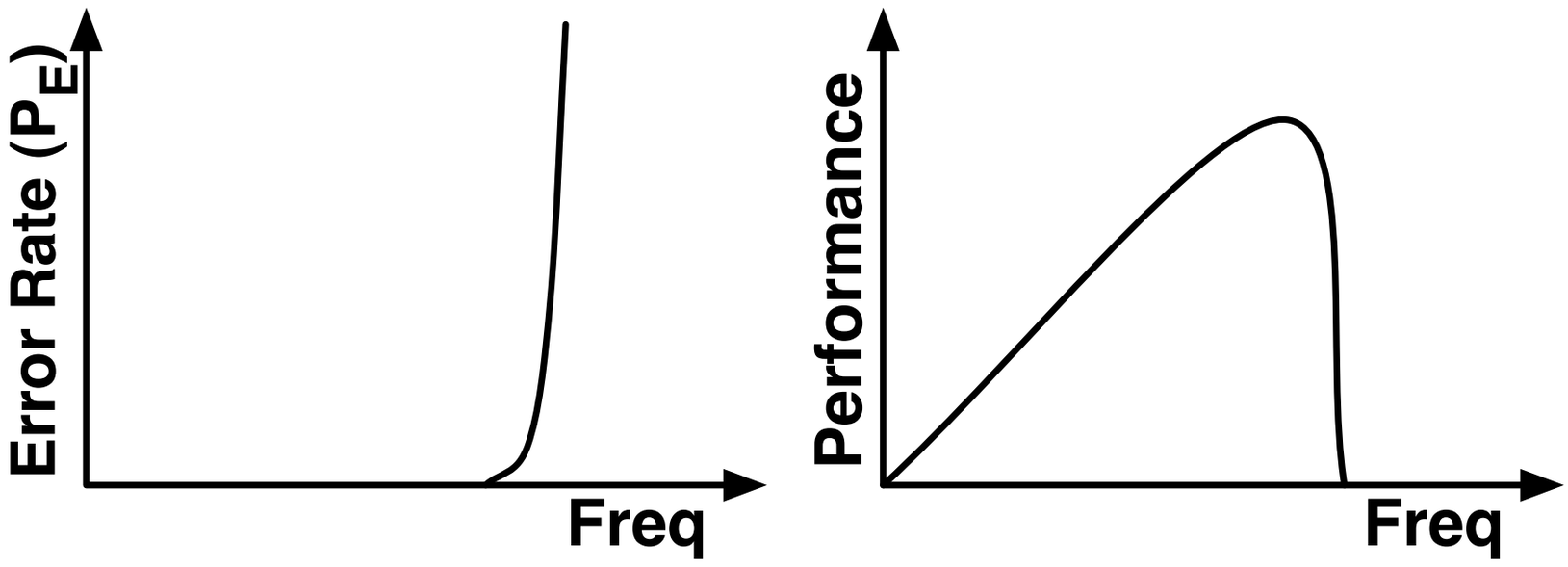
Timing Speculation (TS)

Timing Speculation (TS)

Boost core frequency beyond nominal at **constant** supply voltage

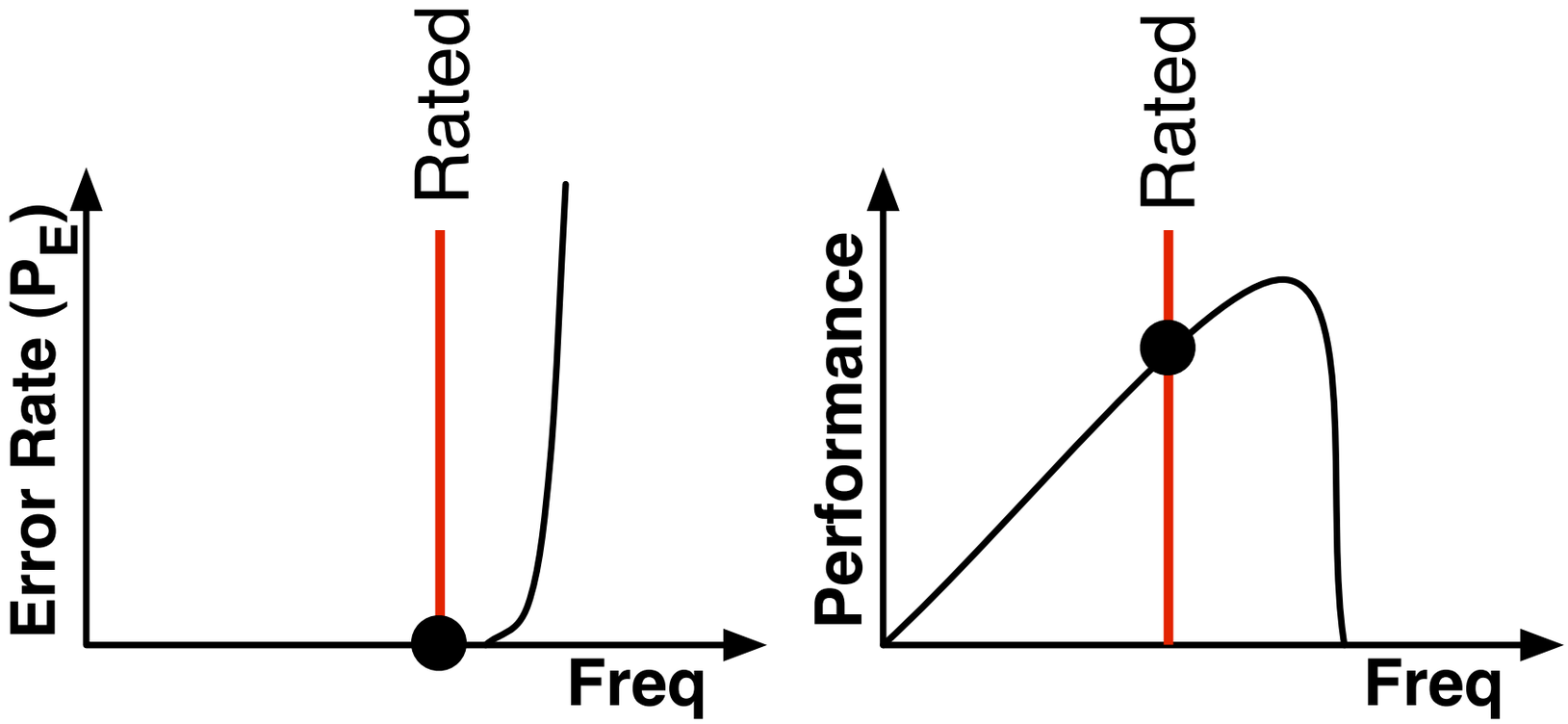
Timing Speculation (TS)

Boost core frequency beyond nominal at **constant** supply voltage



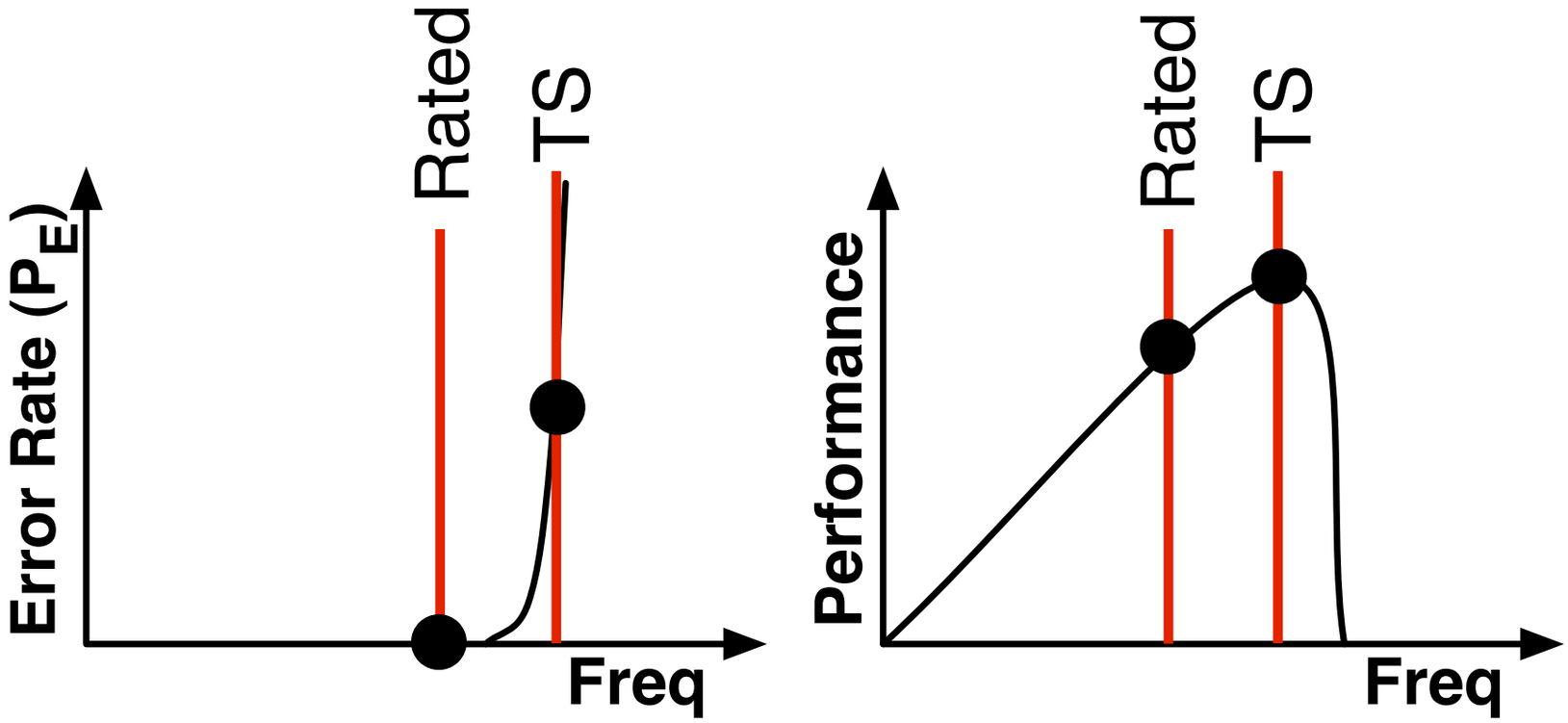
Timing Speculation (TS)

Boost core frequency beyond nominal at **constant** supply voltage



Timing Speculation (TS)

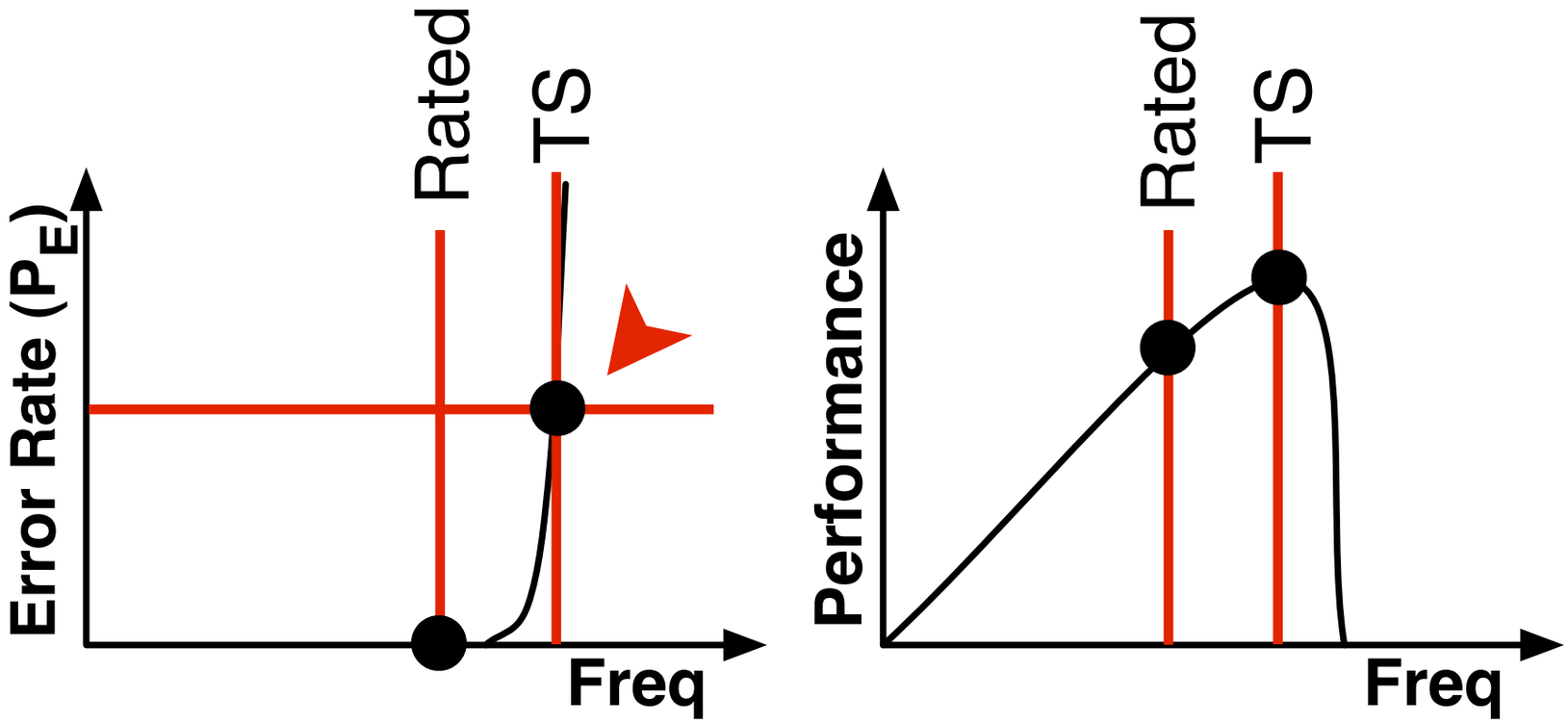
Boost core frequency beyond nominal at **constant** supply voltage



- Increase **f** at constant **V**

Timing Speculation (TS)

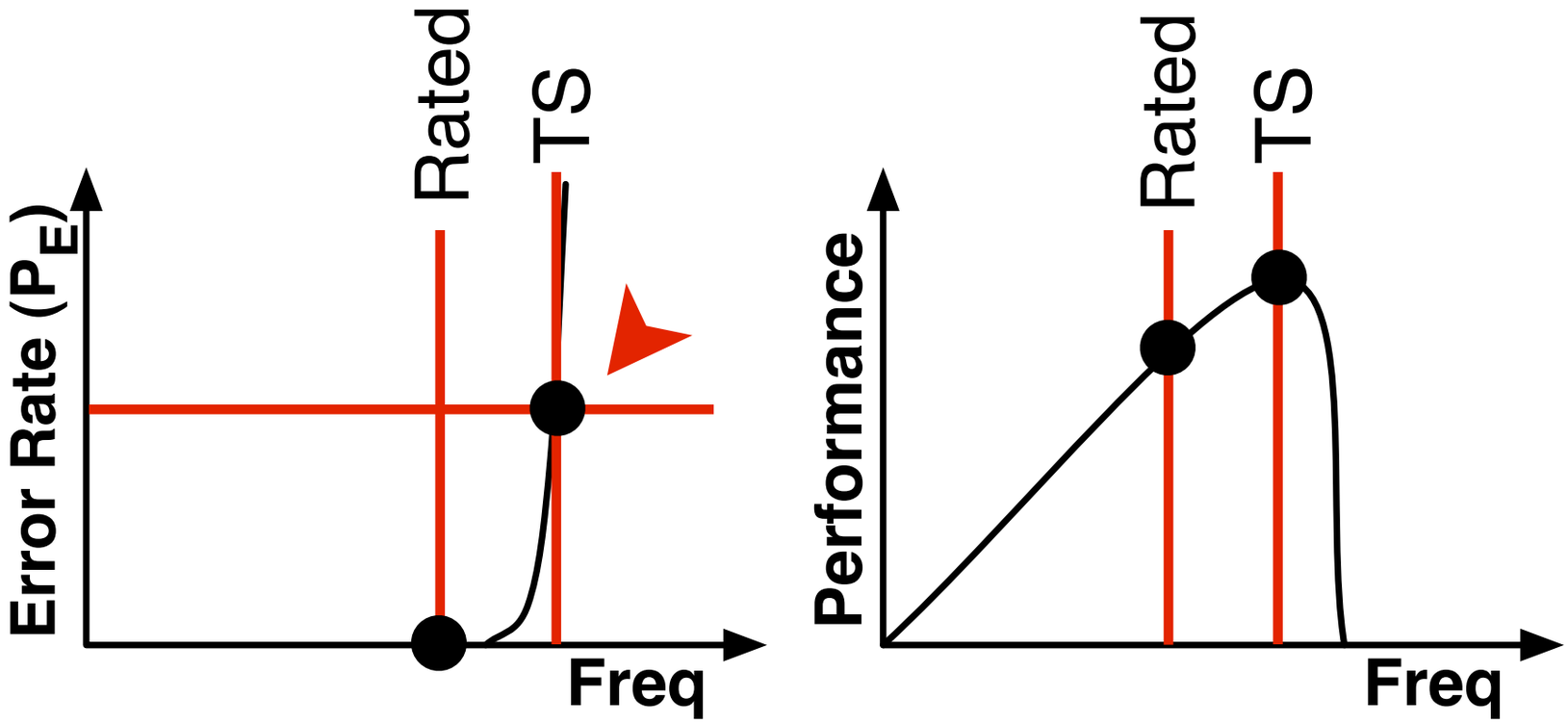
Boost core frequency beyond nominal at **constant** supply voltage



- Increase **f** at constant **V** → Timing errors

Timing Speculation (TS)

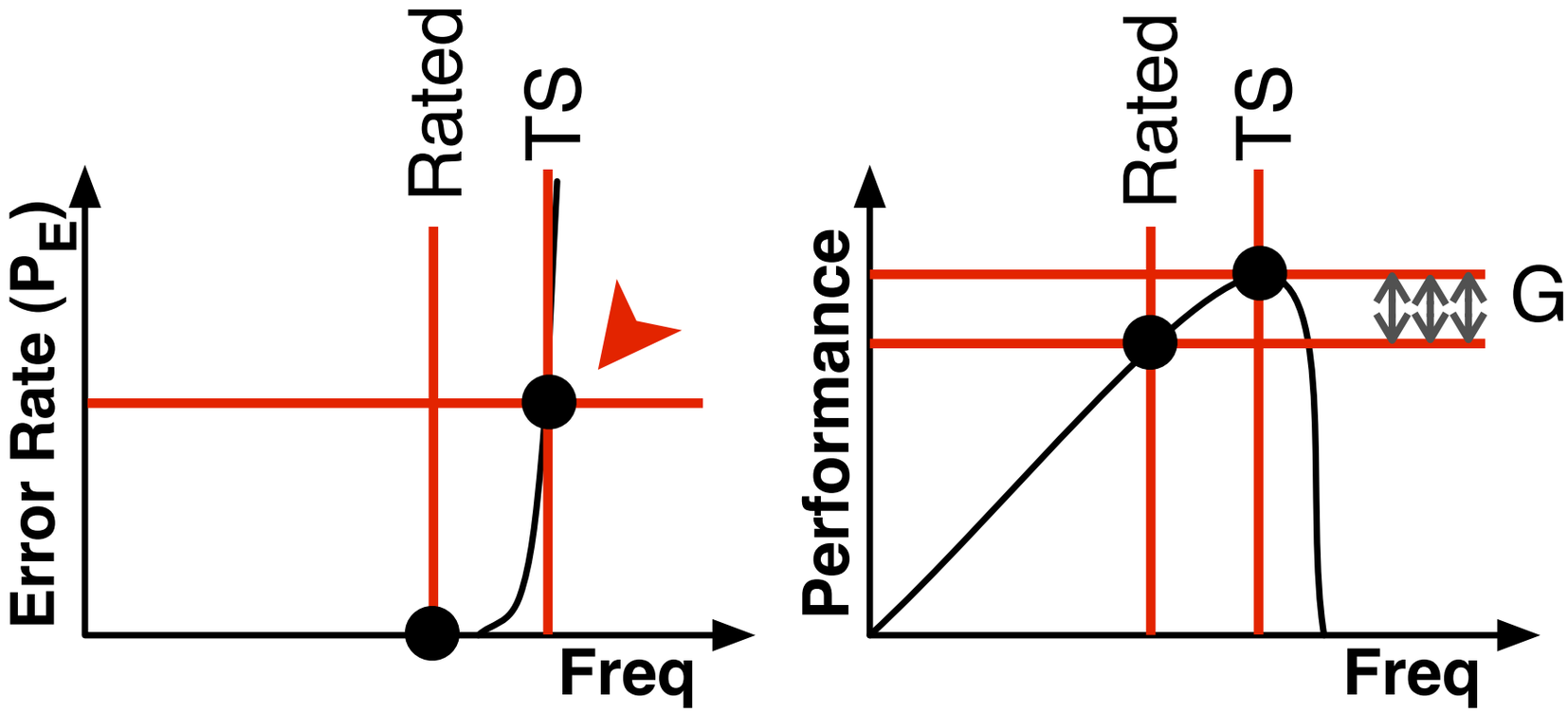
Boost core frequency beyond nominal at **constant** supply voltage



- Increase **f** at constant **V** → Timing errors
- Support for error detection and correction

Timing Speculation (TS)

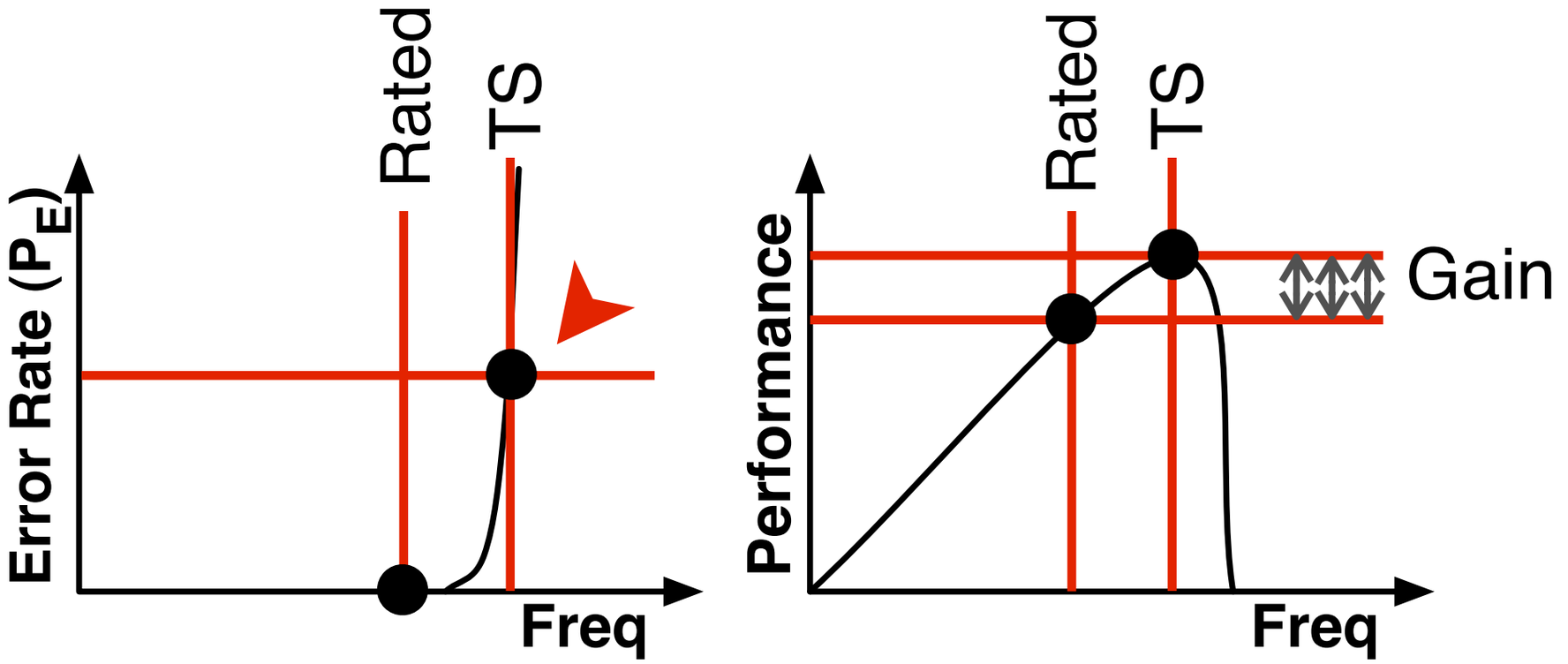
Boost core frequency beyond nominal at **constant** supply voltage



- Increase **f** at constant **V** → Timing errors
- Support for error detection and correction

Timing Speculation (TS)

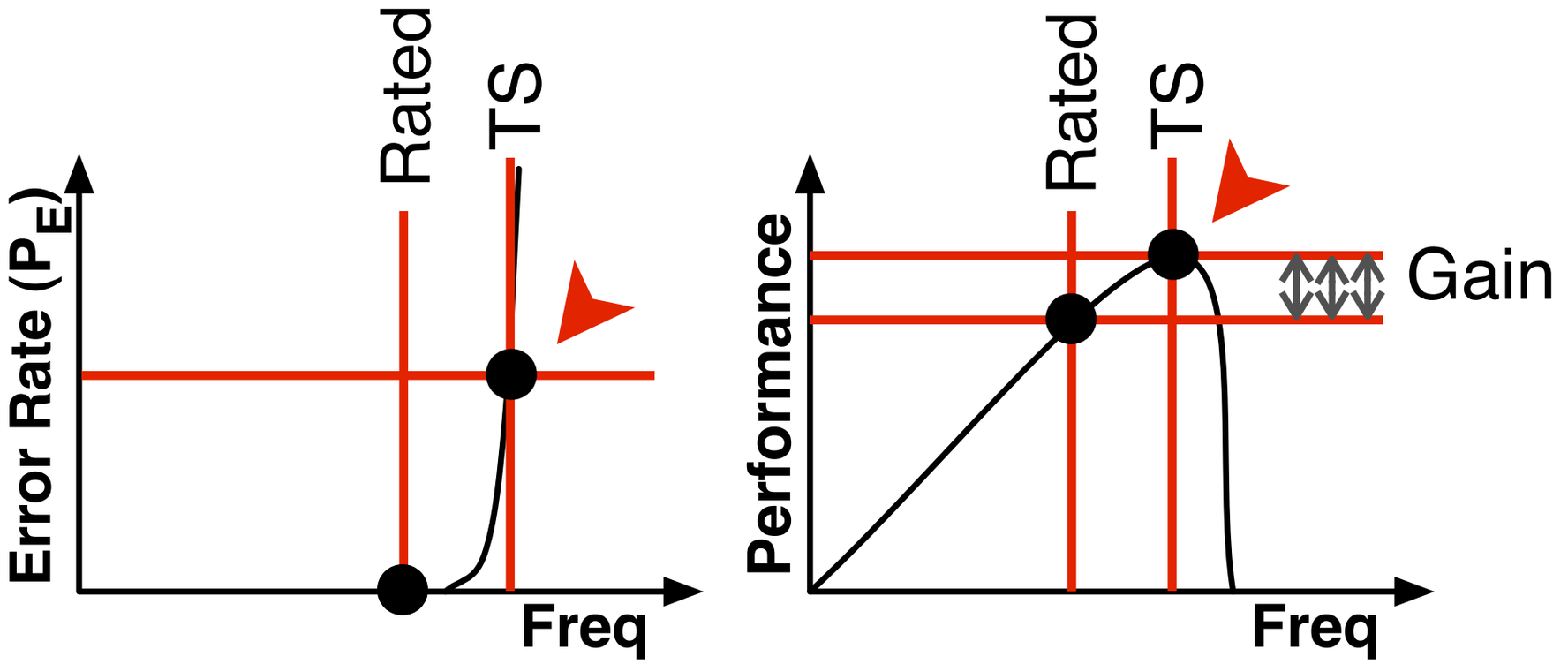
Boost core frequency beyond nominal at **constant** supply voltage



Assuming a high P/T headroom, P_E limits performance gain

Timing Speculation (TS)

Boost core frequency beyond nominal at **constant** supply voltage



Assuming a high P/T headroom, P_E limits performance gain

Voltage-Frequency Boosting

Voltage-Frequency Boosting

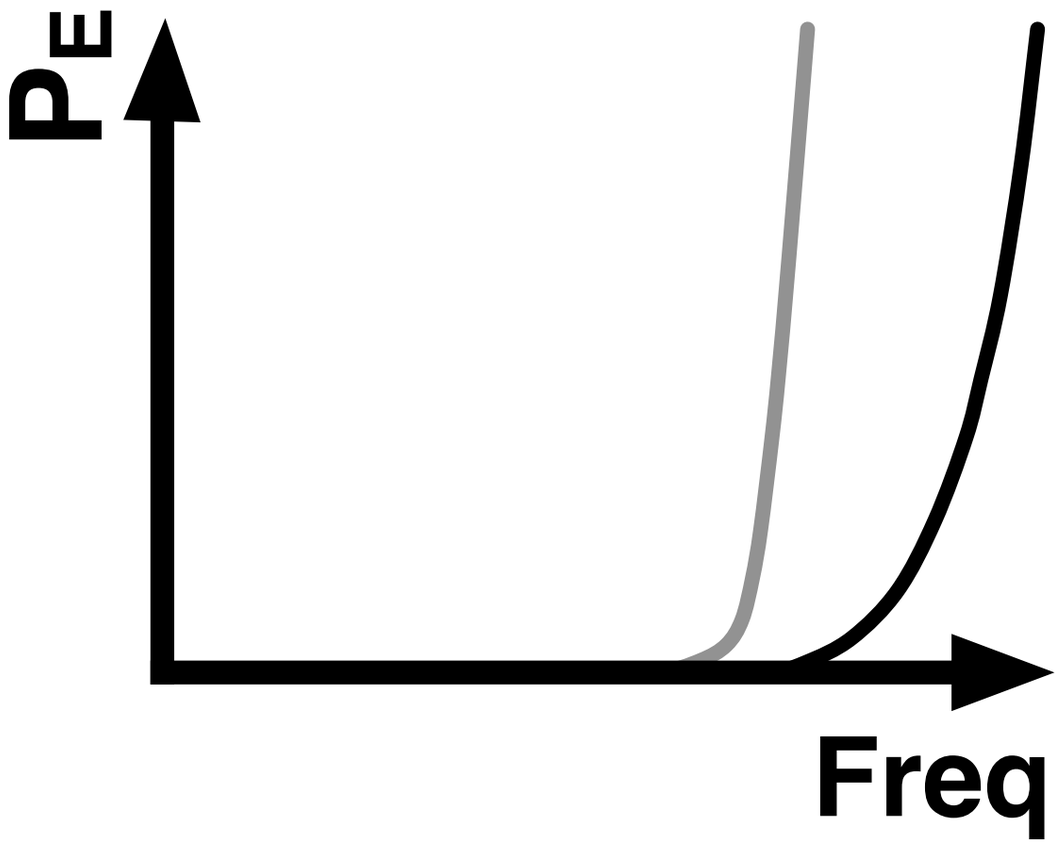
Boost core frequency beyond nominal by **increasing** V

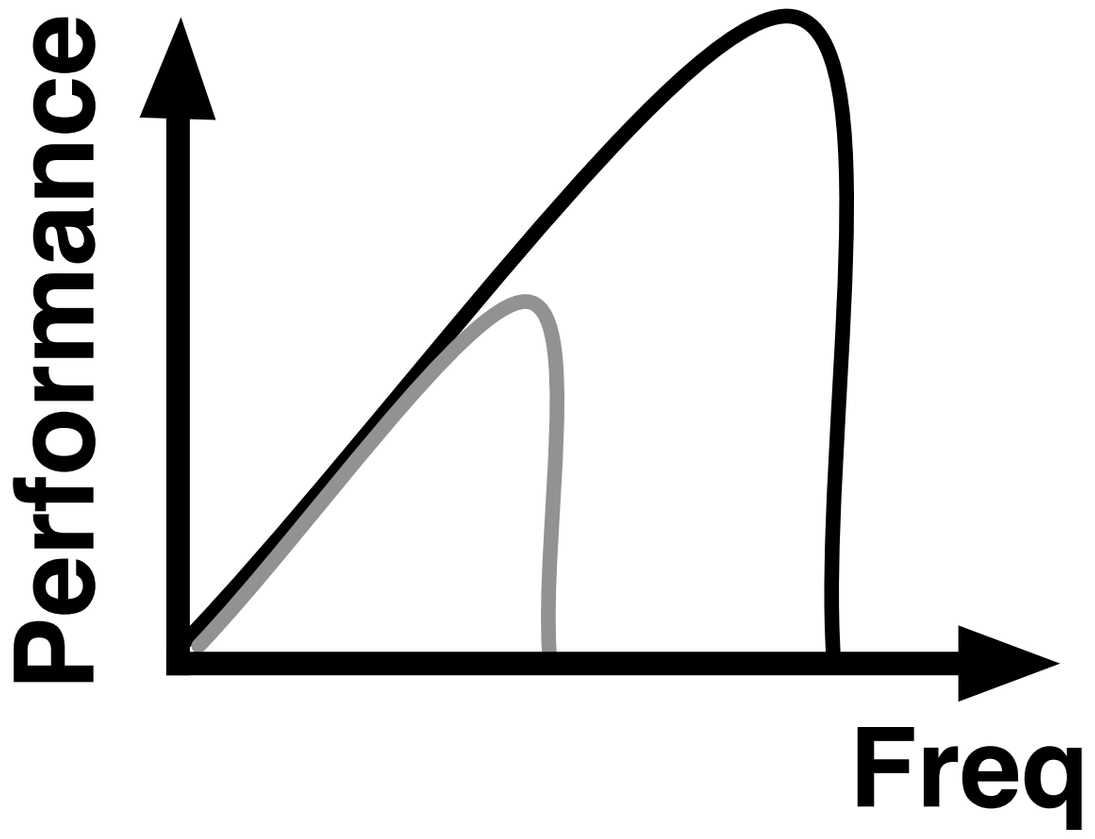
Voltage-Frequency Boosting

Boost core frequency beyond nominal by **increasing** V → No timing errors ($P_E = 0$)

Voltage-Frequency Boosting

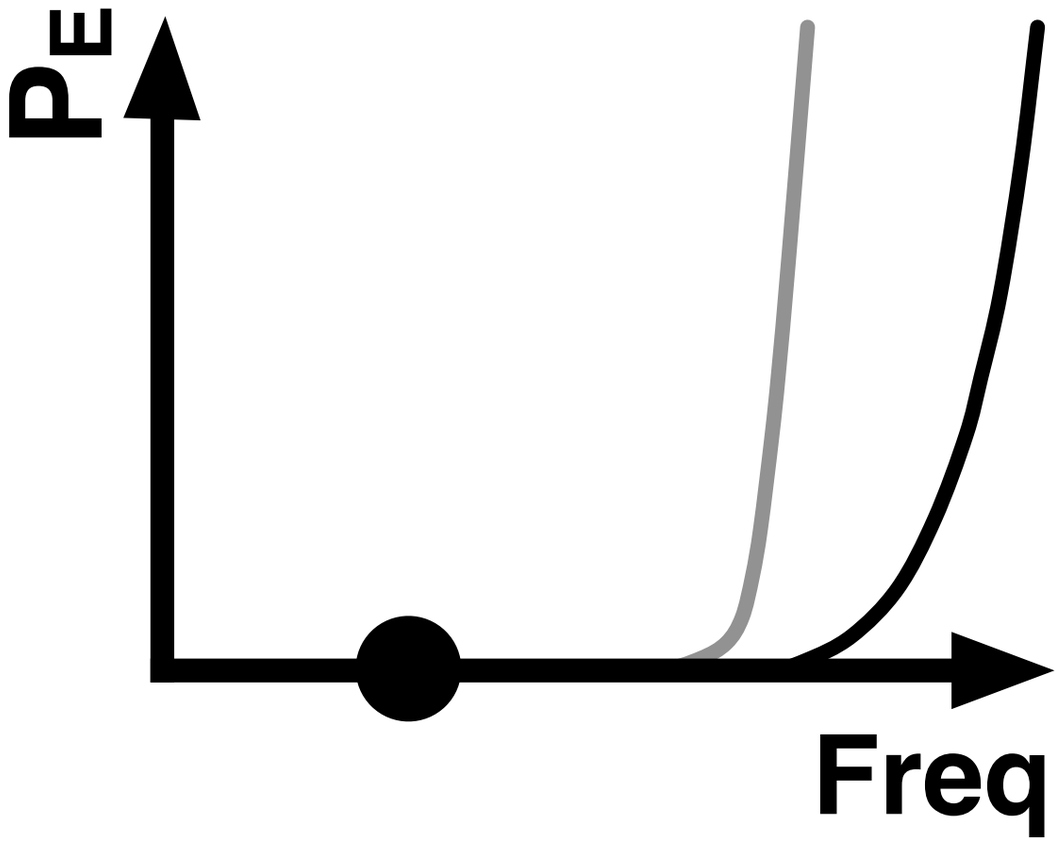
Boost core frequency beyond nominal by **increasing** V → No timing errors ($P_E = 0$)

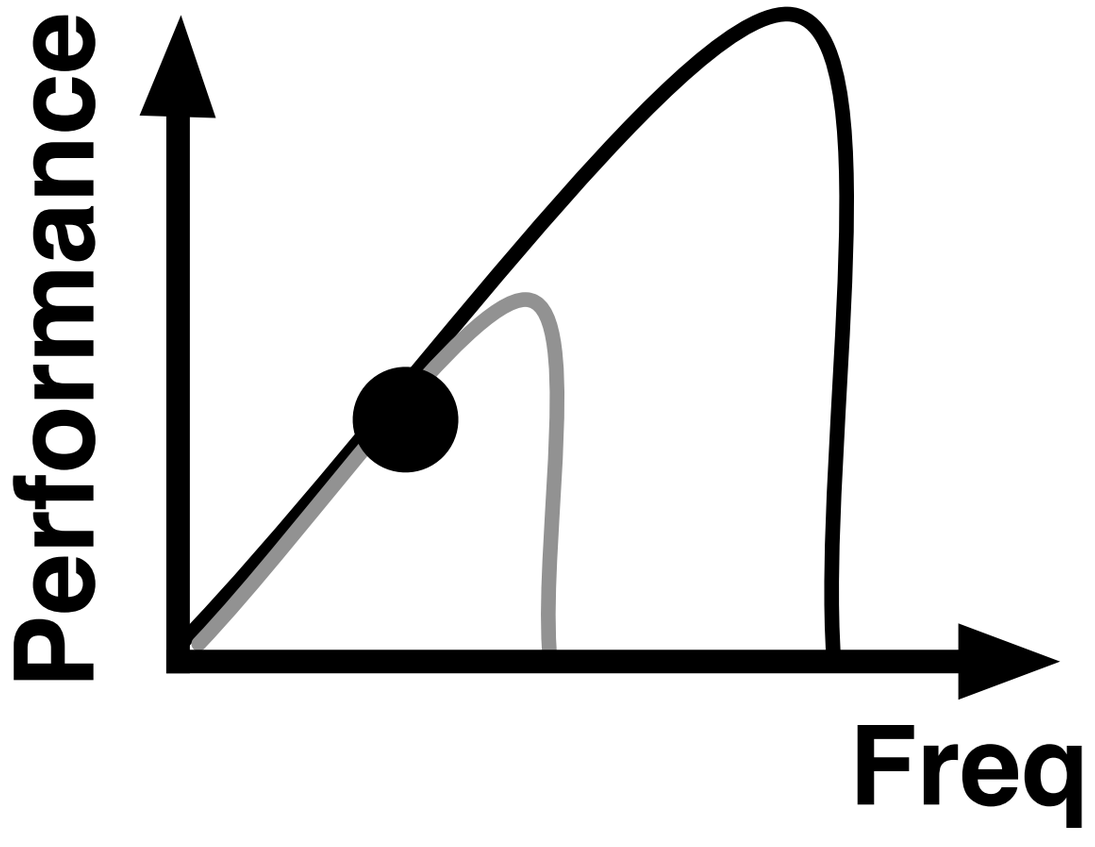




Voltage-Frequency Boosting

Boost core frequency beyond nominal by **increasing** V → No timing errors ($P_E = 0$)

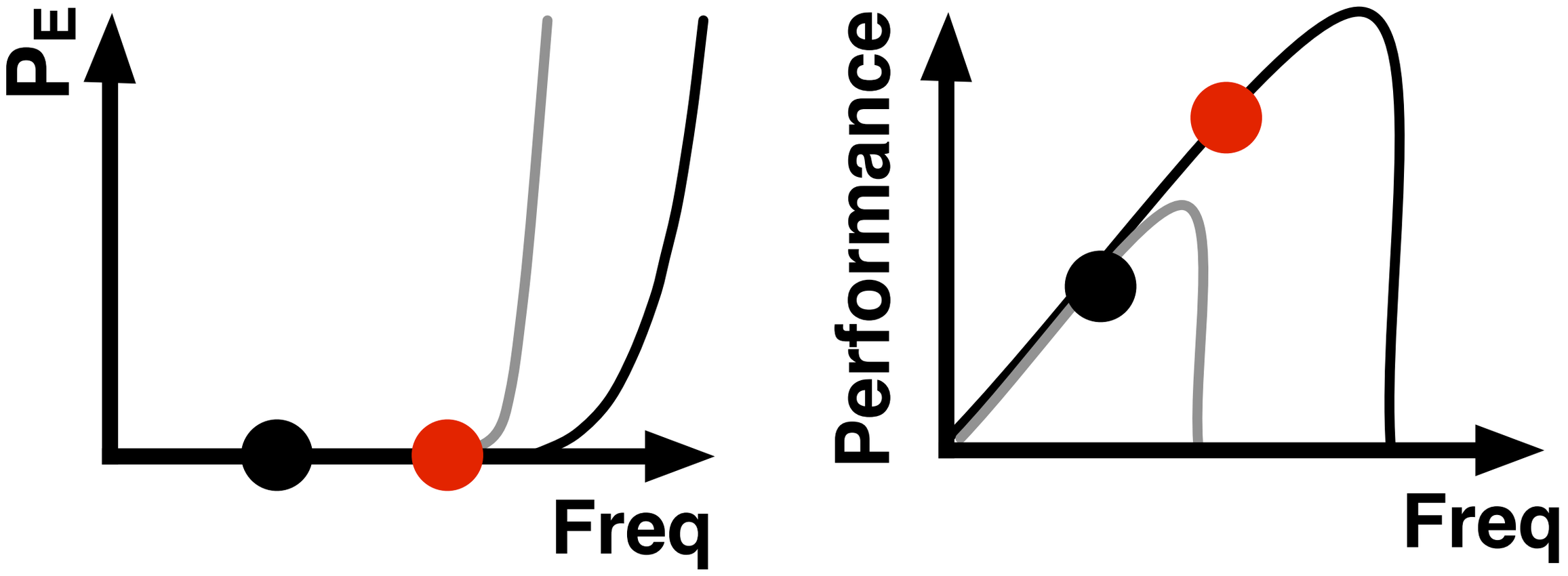




● Rated

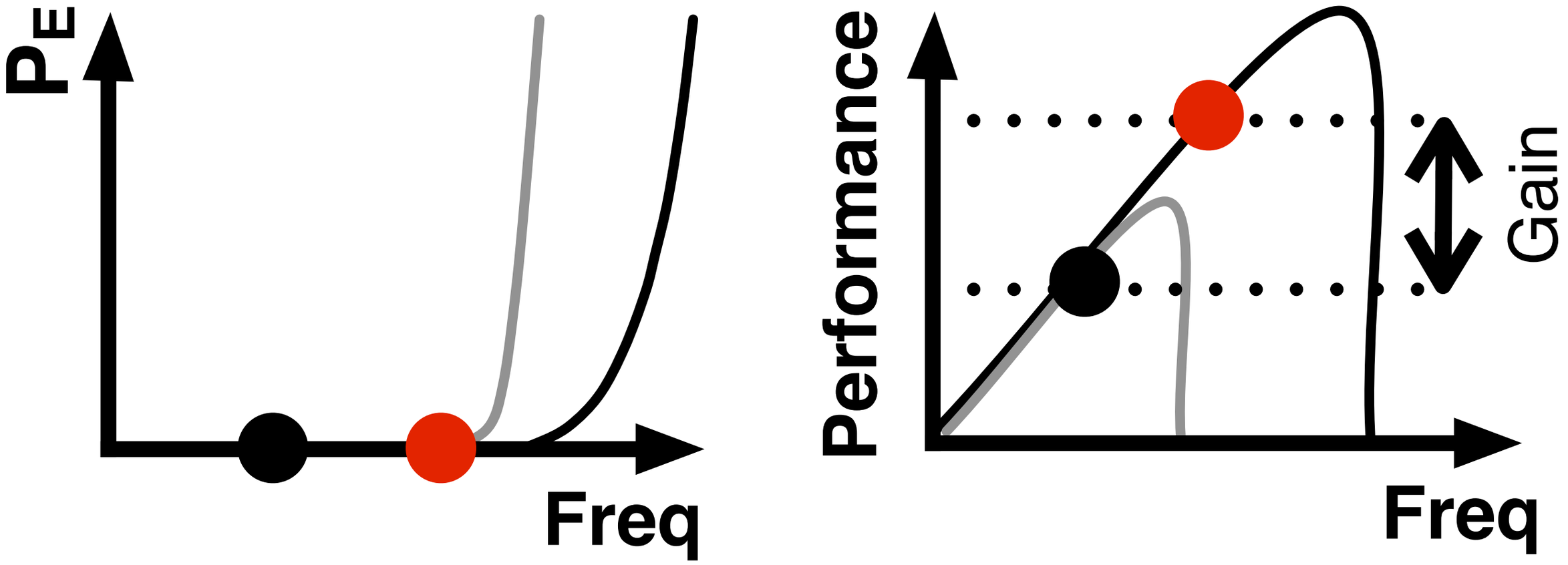
Voltage-Frequency Boosting

Boost core frequency beyond nominal by **increasing** V → No timing errors ($P_E = 0$)



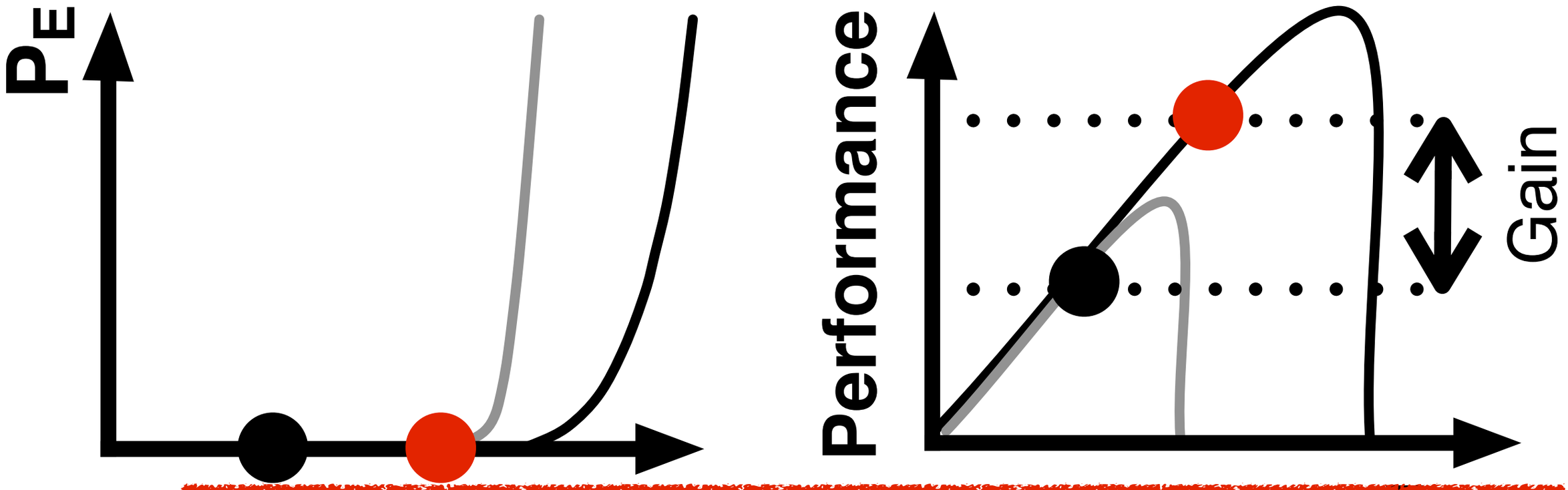
Voltage-Frequency Boosting

Boost core frequency beyond nominal by
increasing V → No timing errors ($P_E = 0$)



Voltage-Frequency Boosting

Boost core frequency beyond nominal by
increasing V → No timing errors ($P_E = 0$)



Assuming a high P/T headroom,
V limits performance gain

Contribution: LeadOut

Contribution: LeadOut

- Individual application of TS or V/f Boosting is suboptimal

Contribution: LeadOut

- Individual application of TS or V/f Boosting is suboptimal
  - Unable to bring the multicore up to its P/T envelope

Contribution: LeadOut

- Individual application of TS or V/f Boosting is suboptimal
  - Unable to bring the multicore up to its P/T envelope
  - Available P/T headroom remains untapped

Contribution: LeadOut

- Individual application of TS or V/f Boosting is suboptimal
  - Unable to bring the multicore up to its P/T envelope
  - Available P/T headroom remains untapped
- TS and V/f boosting are complementary

Contribution: LeadOut

- Individual application of TS or V/f Boosting is suboptimal
  - Unable to bring the multicore up to its P/T envelope
  - Available P/T headroom remains untapped
- TS and V/f boosting are complementary
  - Bounded by different constraints

Contribution: LeadOut

- Individual application of TS or V/f Boosting is suboptimal
  - Unable to bring the multicore up to its P/T envelope
  - Available P/T headroom remains untapped
- TS and V/f boosting are complementary
  - Bounded by different constraints
- LeadOut

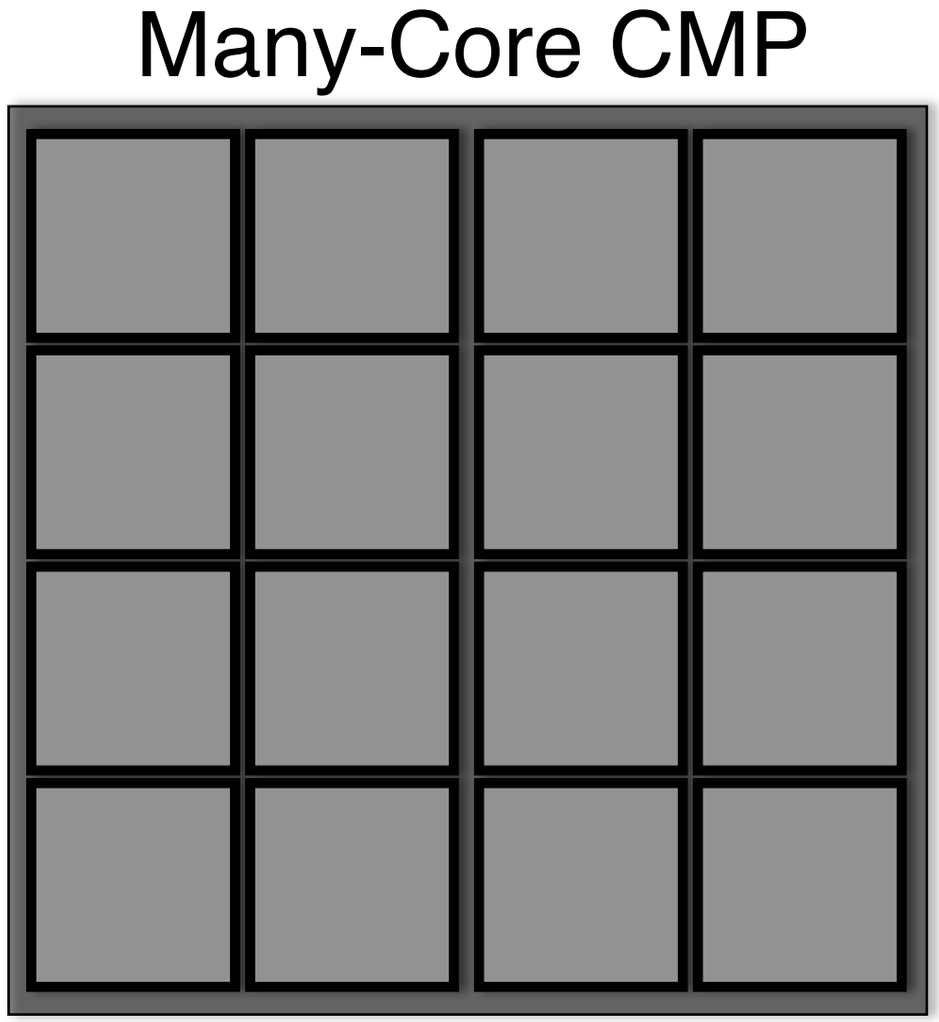
Contribution: LeadOut

- Individual application of TS or V/f Boosting is suboptimal
  - Unable to bring the multicore up to its P/T envelope
  - Available P/T headroom remains untapped
- TS and V/f boosting are complementary
  - Bounded by different constraints
- LeadOut
  - Synergistically combine TS and V/f Boosting

Contribution: LeadOut

- Individual application of TS or V/f Boosting is suboptimal
  - Unable to bring the multicore up to its P/T envelope
  - Available P/T headroom remains untapped
- TS and V/f boosting are complementary
  - Bounded by different constraints
- LeadOut
  - Synergistically combine TS and V/f Boosting
  - Speed-ups for single thread performance multiply

Example TS Architecture [PACT07]: Paceline

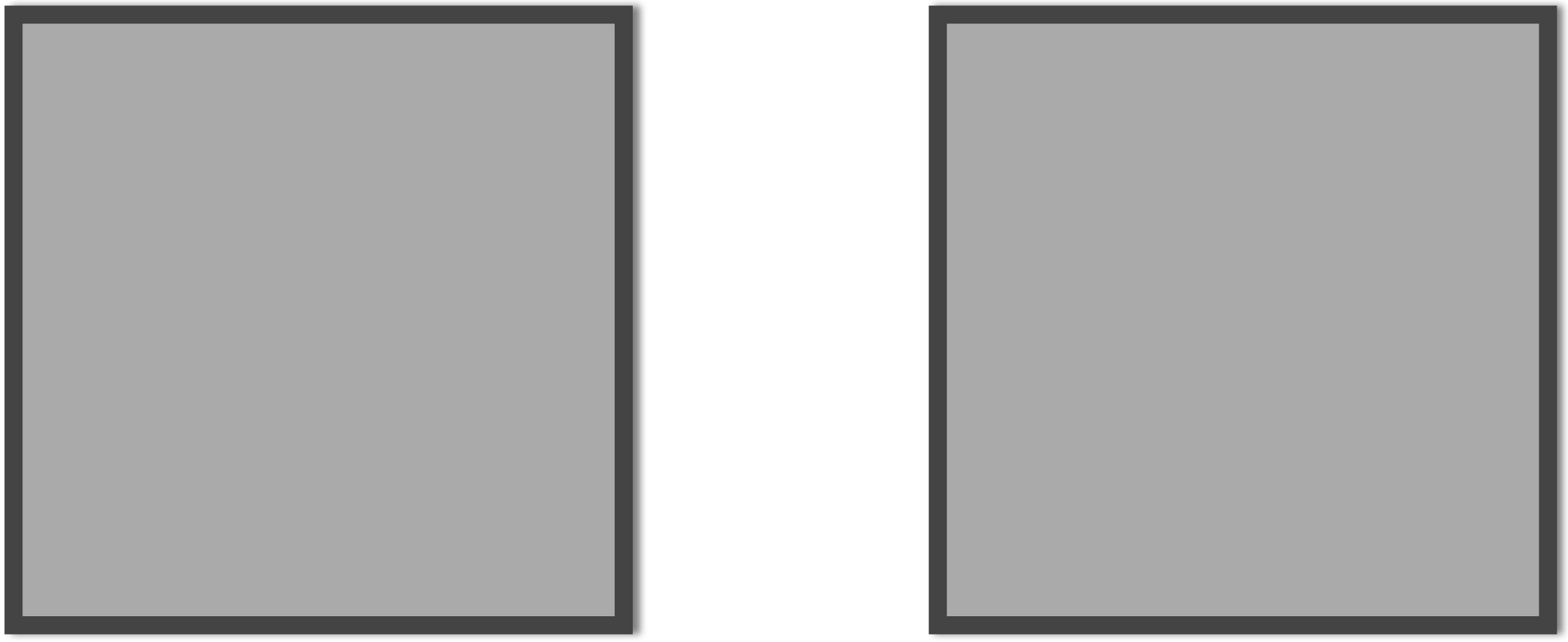


Example TS Architecture [PACT07]: Paceline

OS sees: One core

Leader

Checker



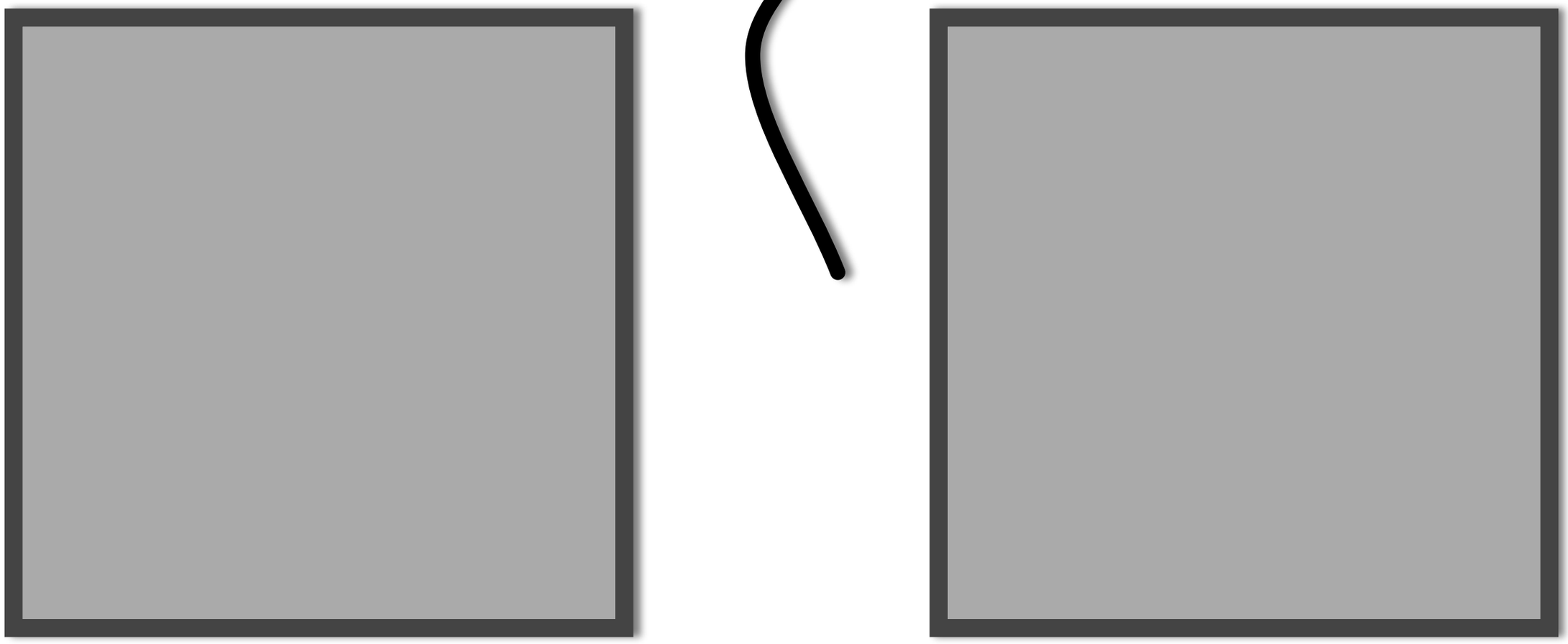
Example TS Architecture [PACT07]: Paceline

OS sees: One core

Critical Thread

Leader

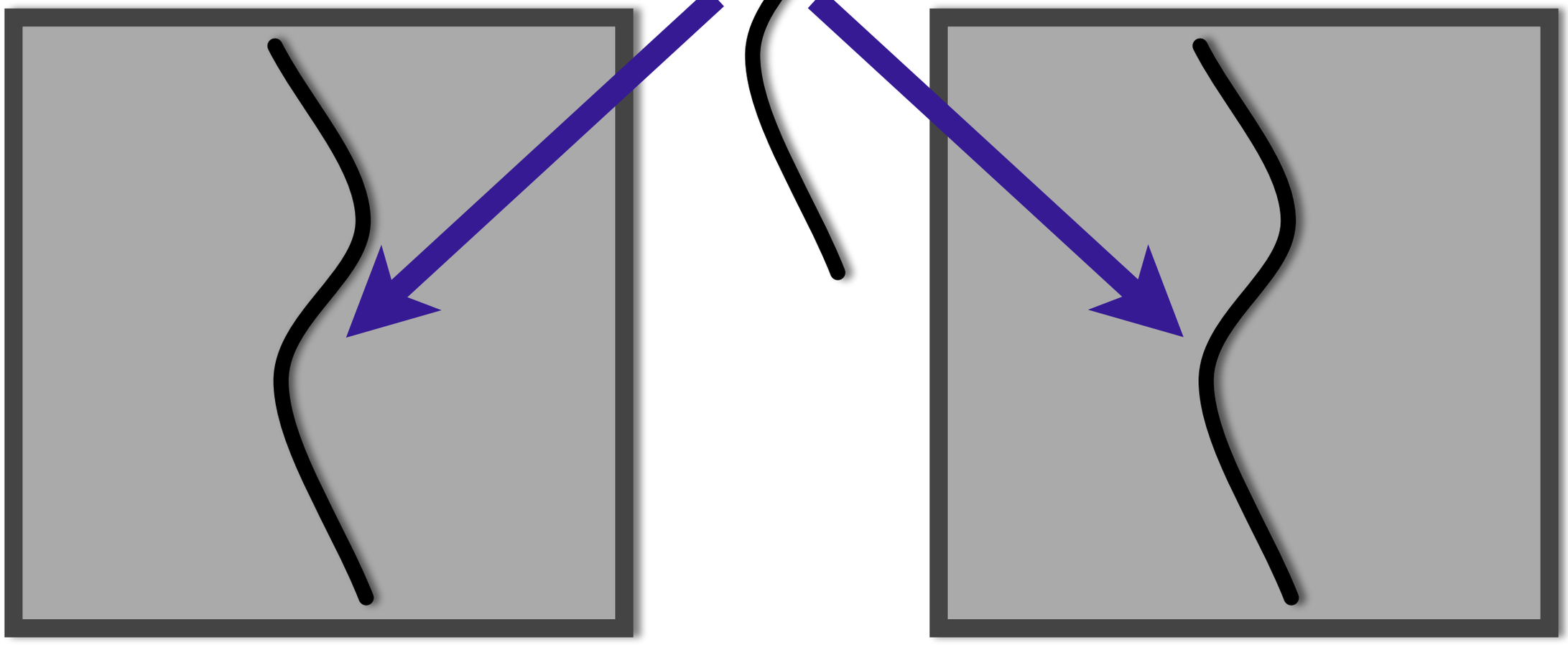
Checker



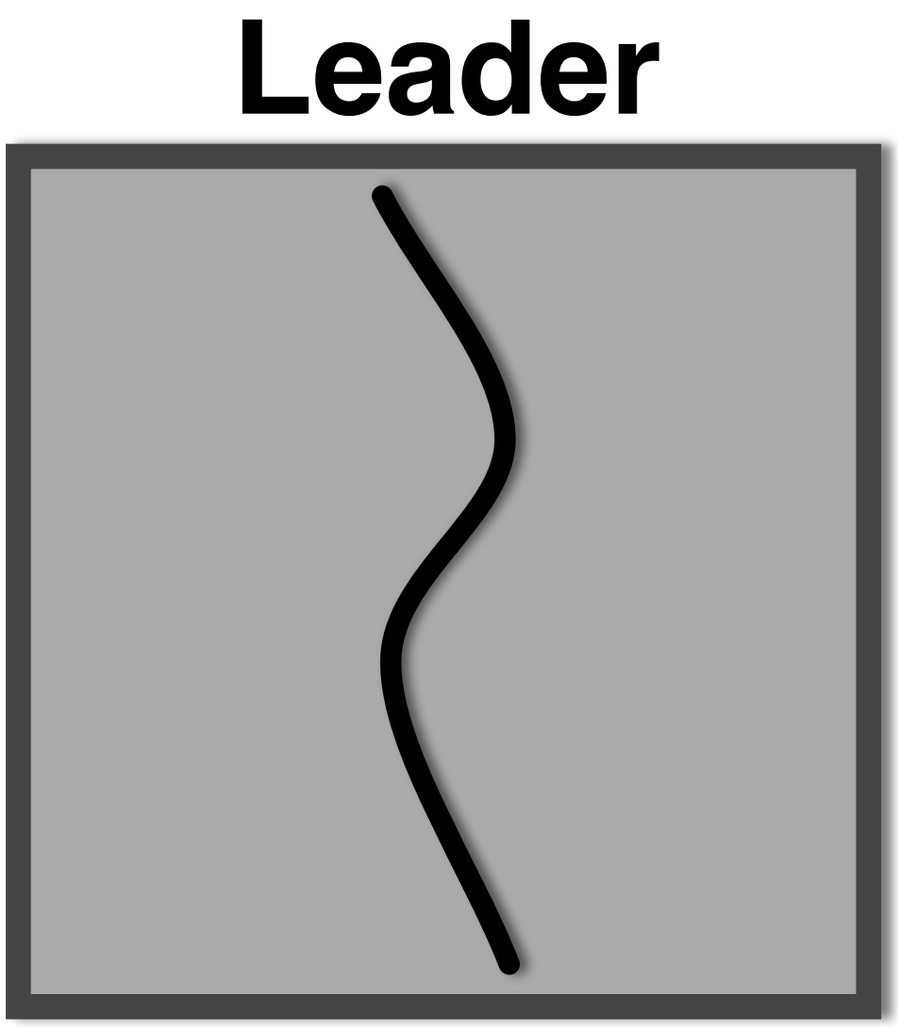
Example TS Architecture [PACT07]: Paceline

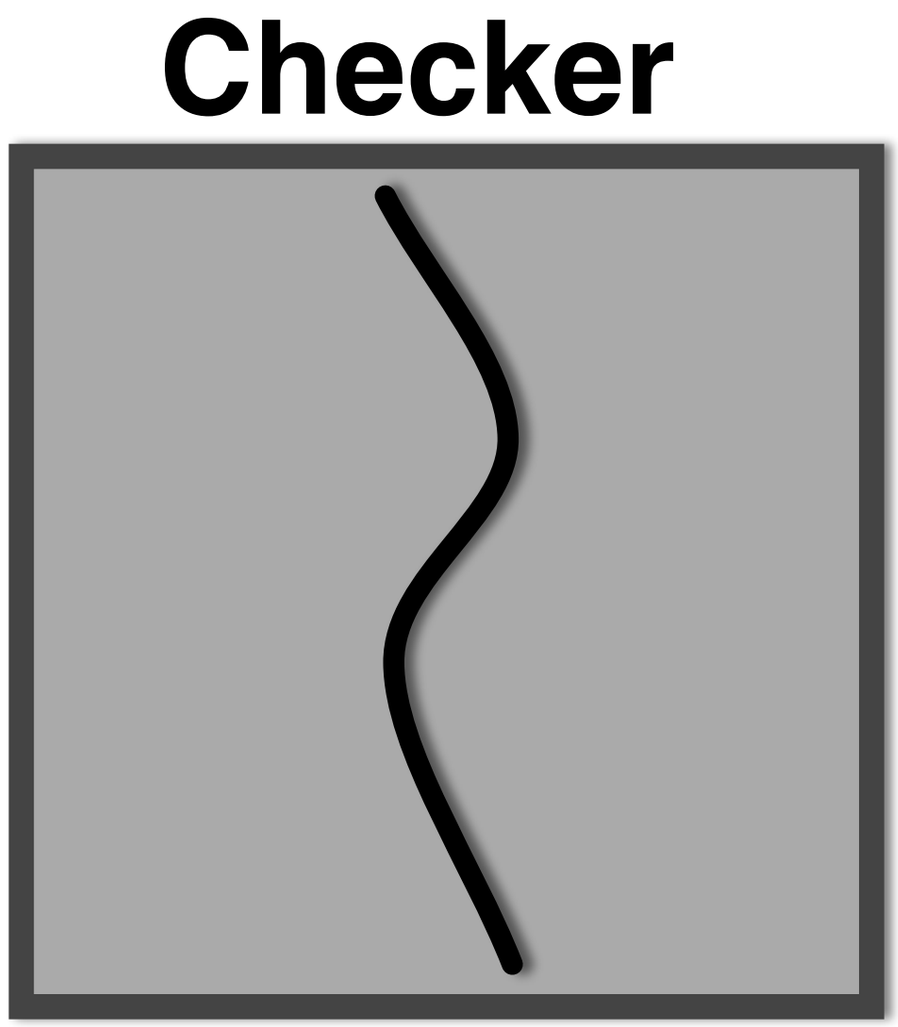
OS sees: One core

Critical Thread

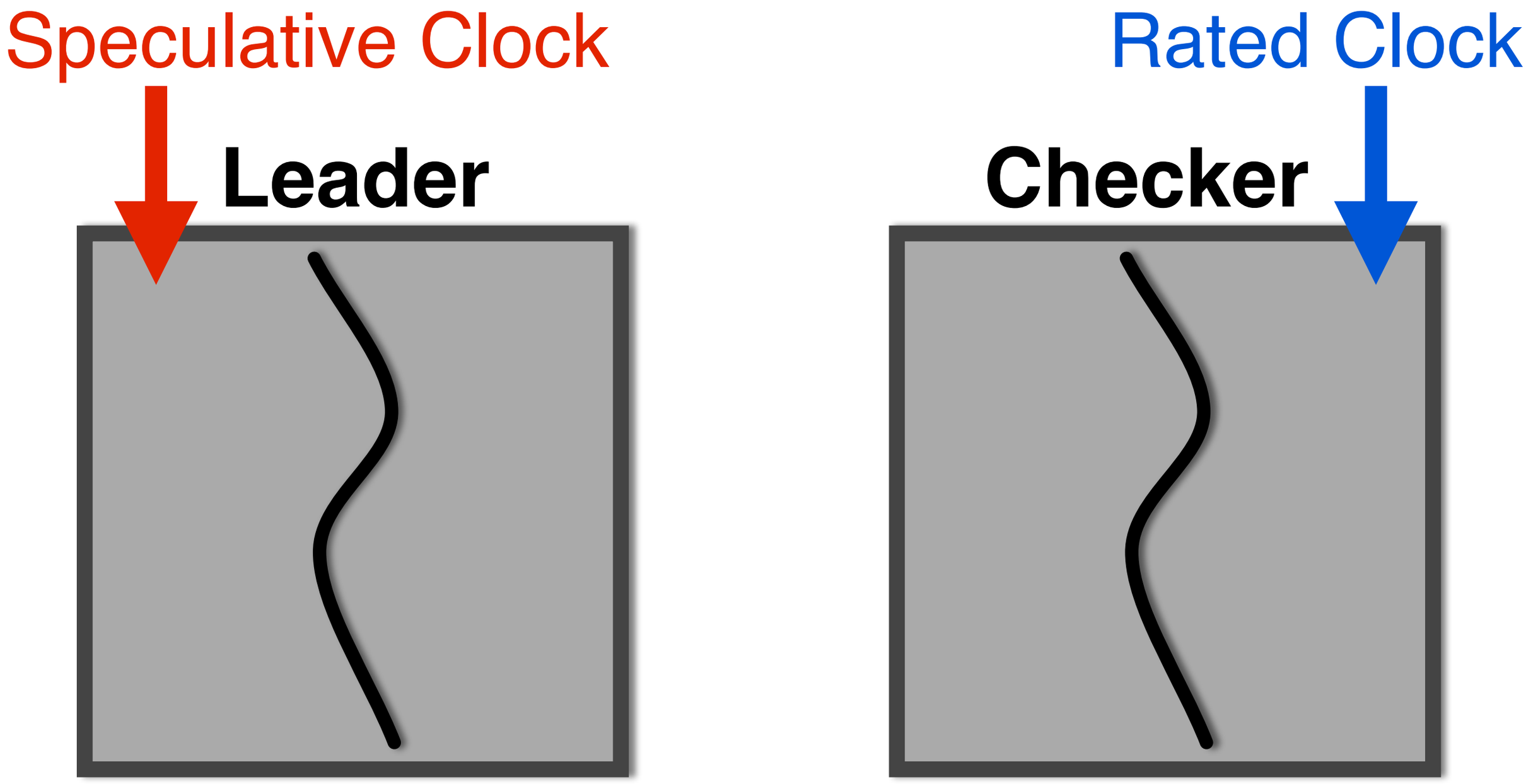


Example TS Architecture [PACT07]: Paceline

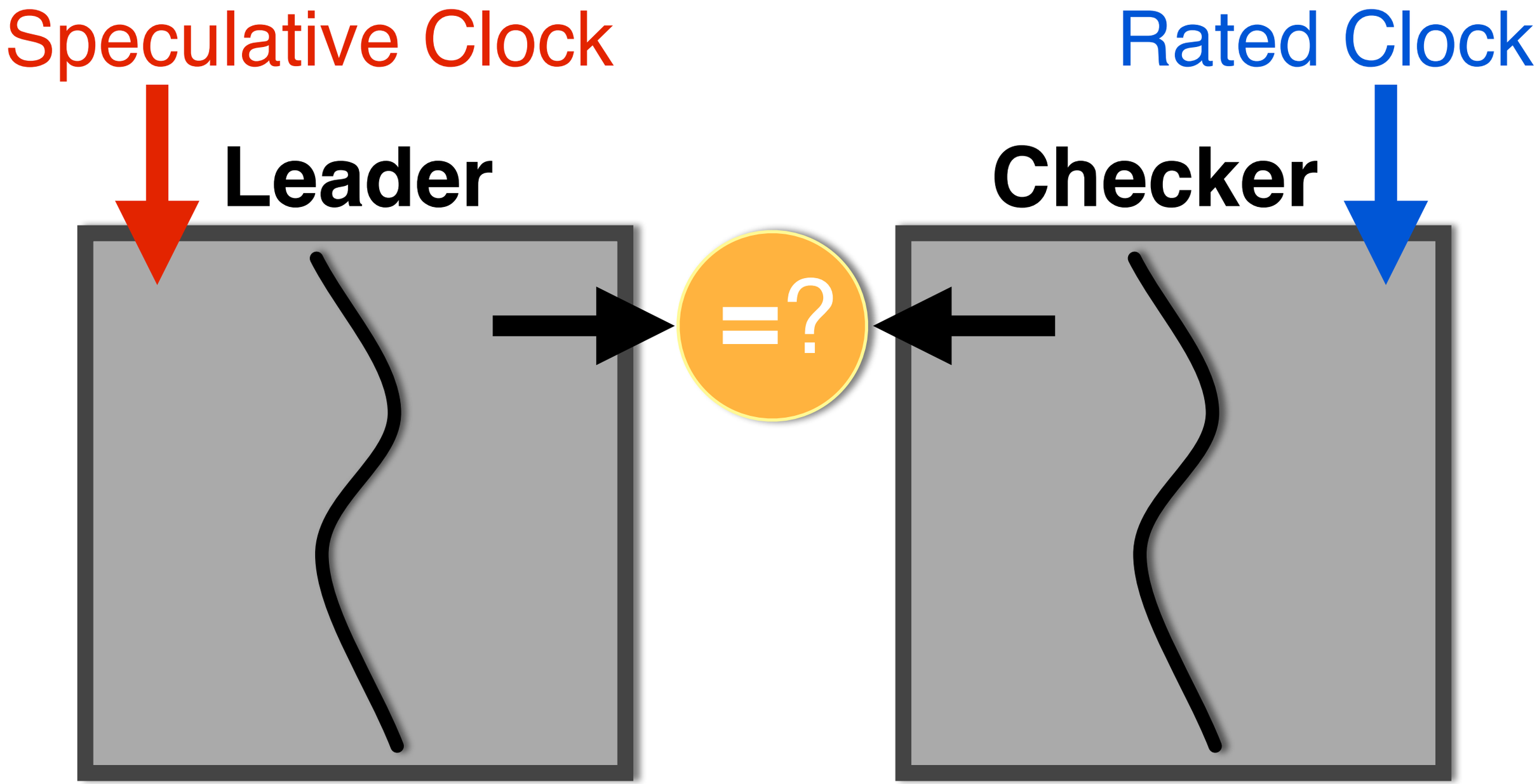




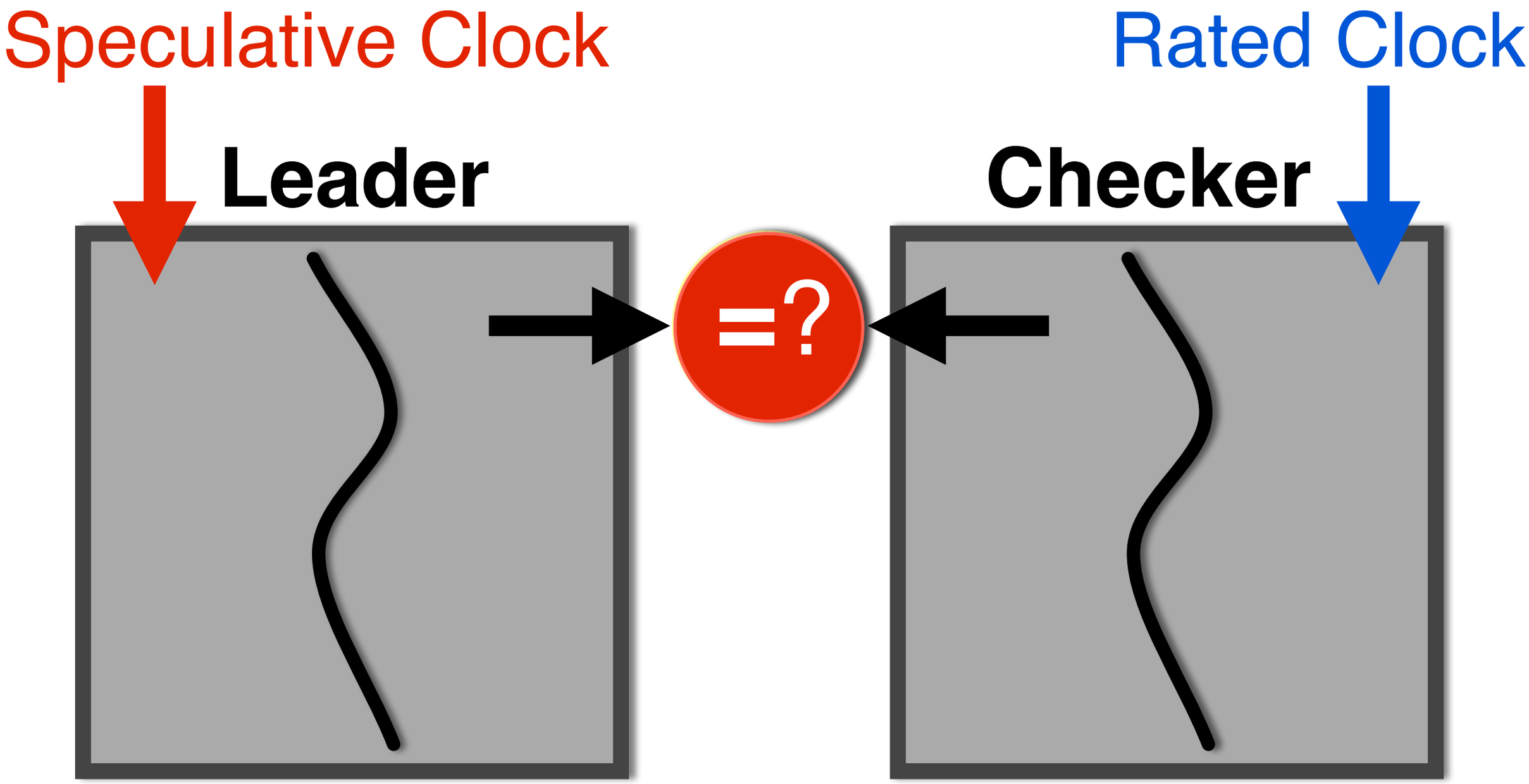
Example TS Architecture [PACT07]: Paceline



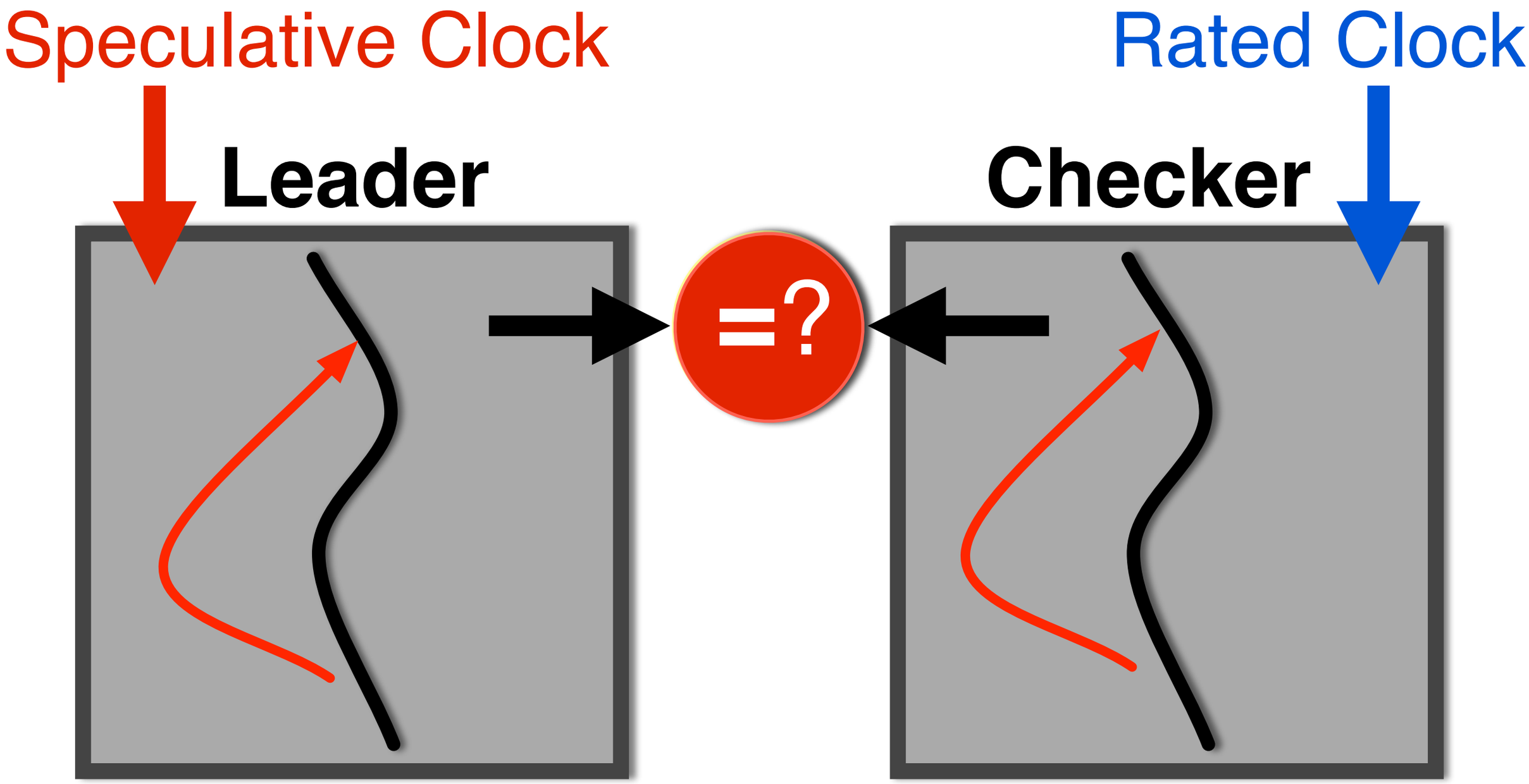
Example TS Architecture [PACT07]: Paceline



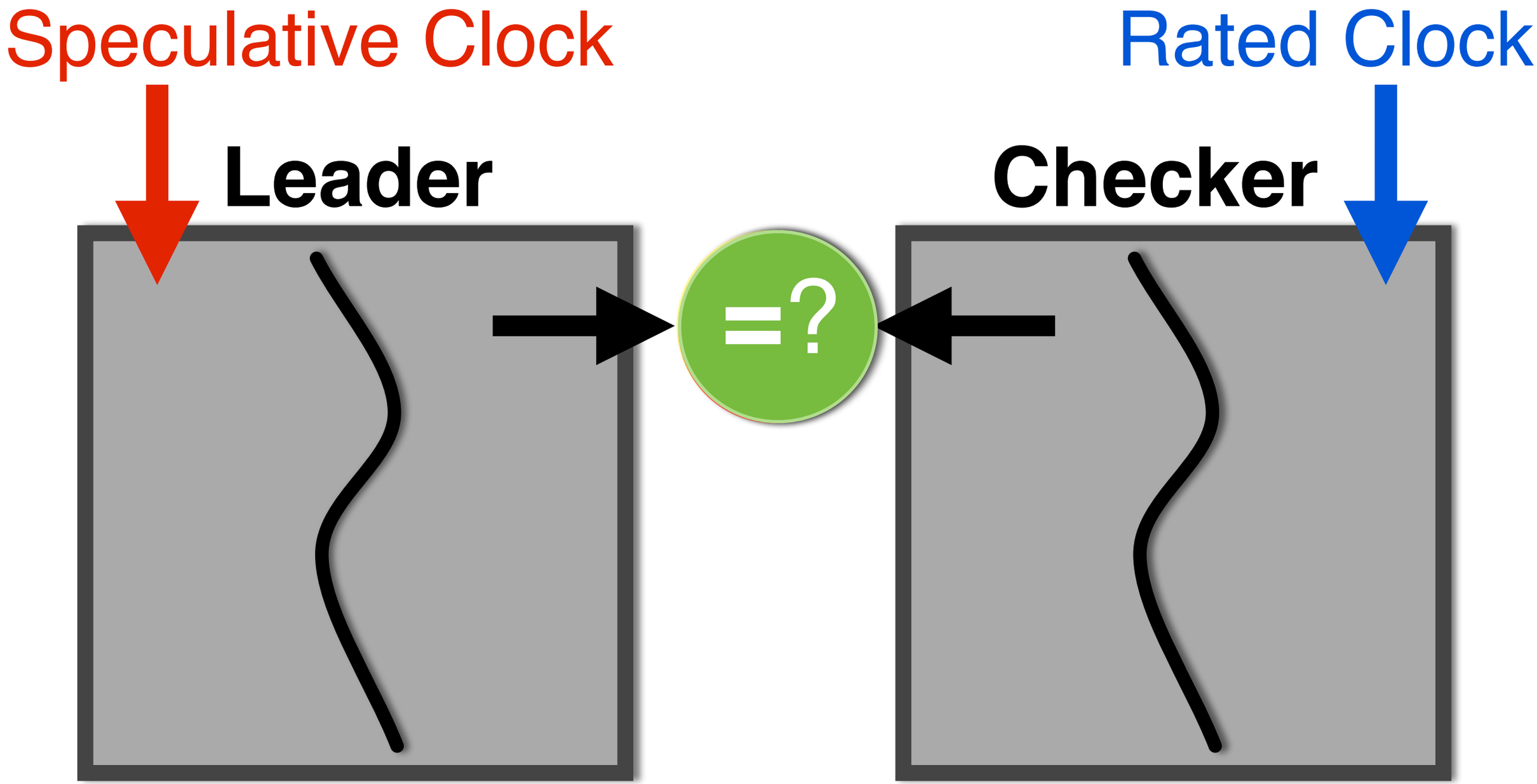
Example TS Architecture [PACT07]: Paceline



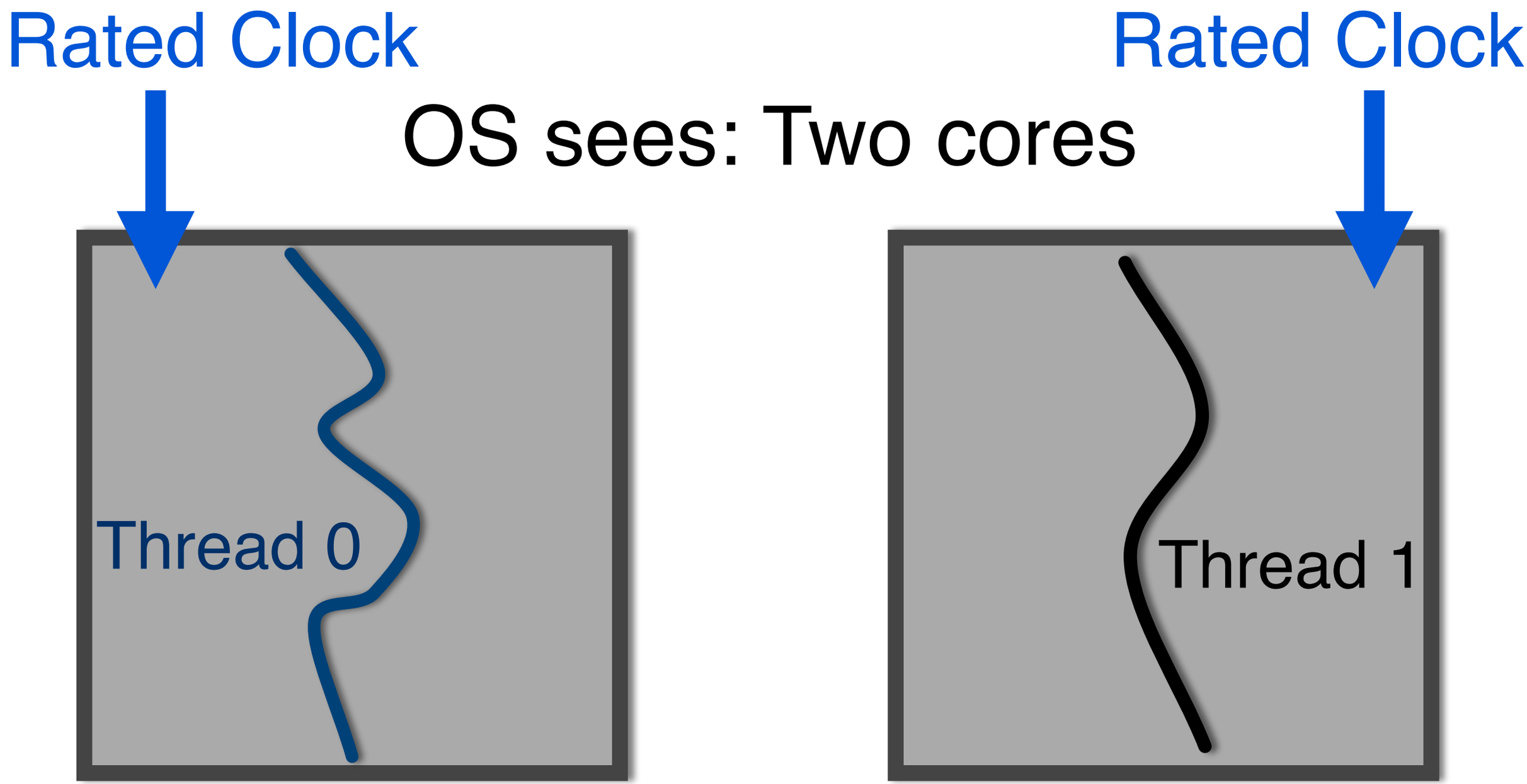
Example TS Architecture [PACT07]: Paceline



Example TS Architecture [PACT07]: Paceline

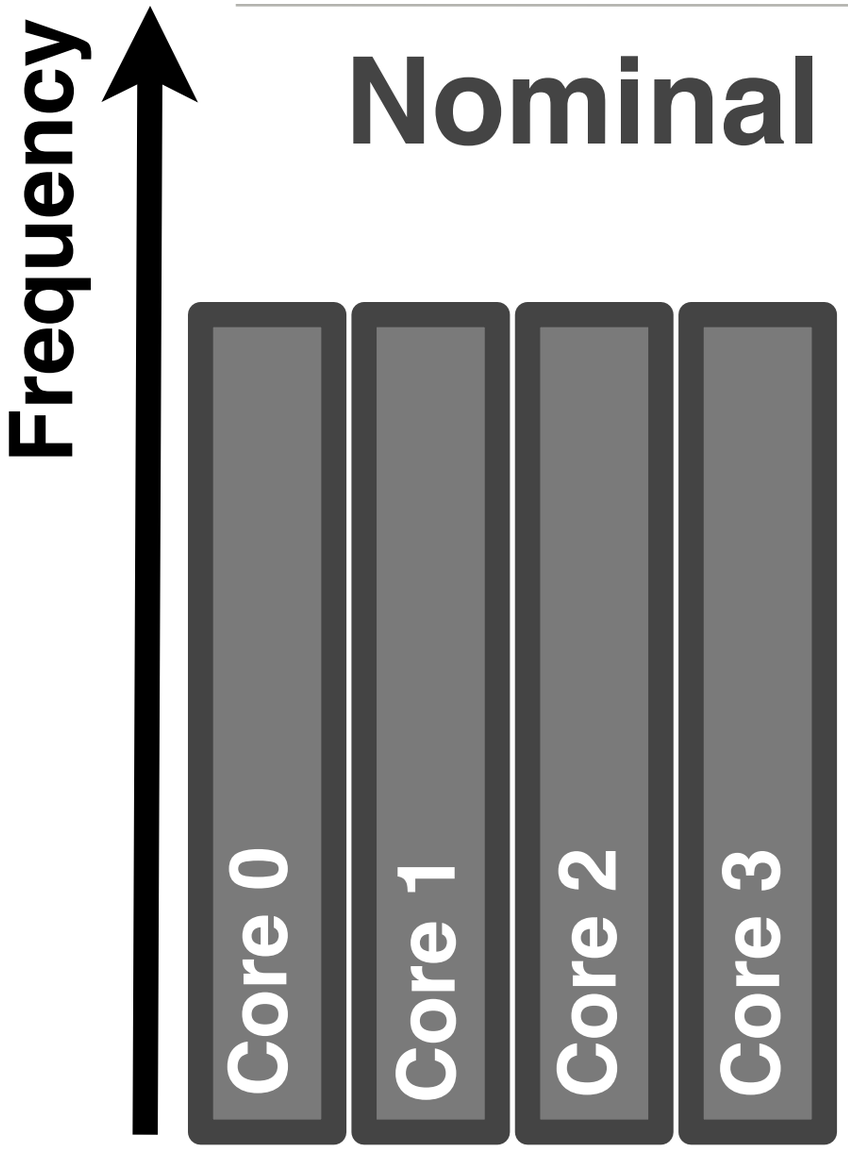


Example TS Architecture [PACT07]: Paceline

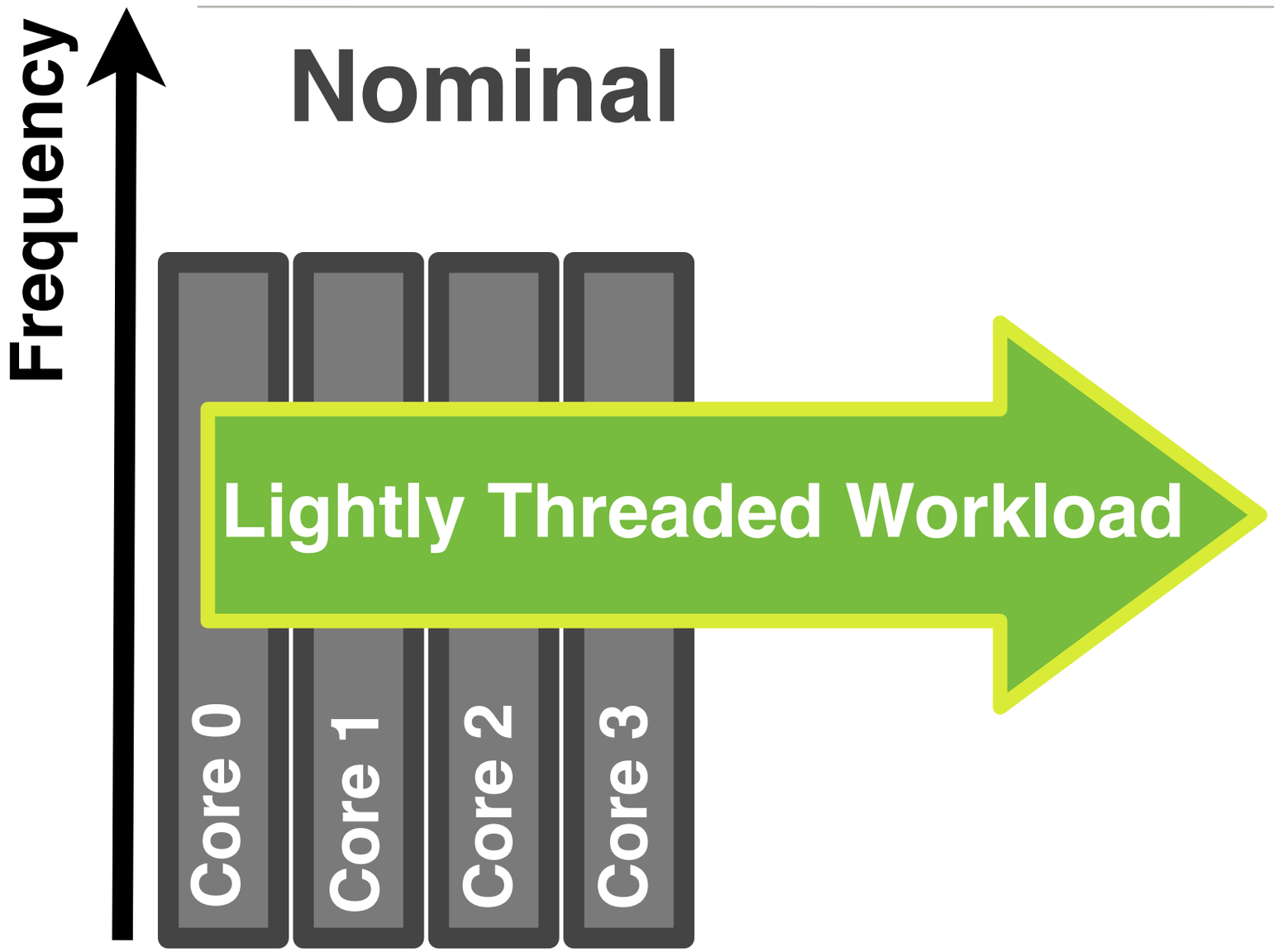


Example V/f Boosting: Intel Turbo Boost

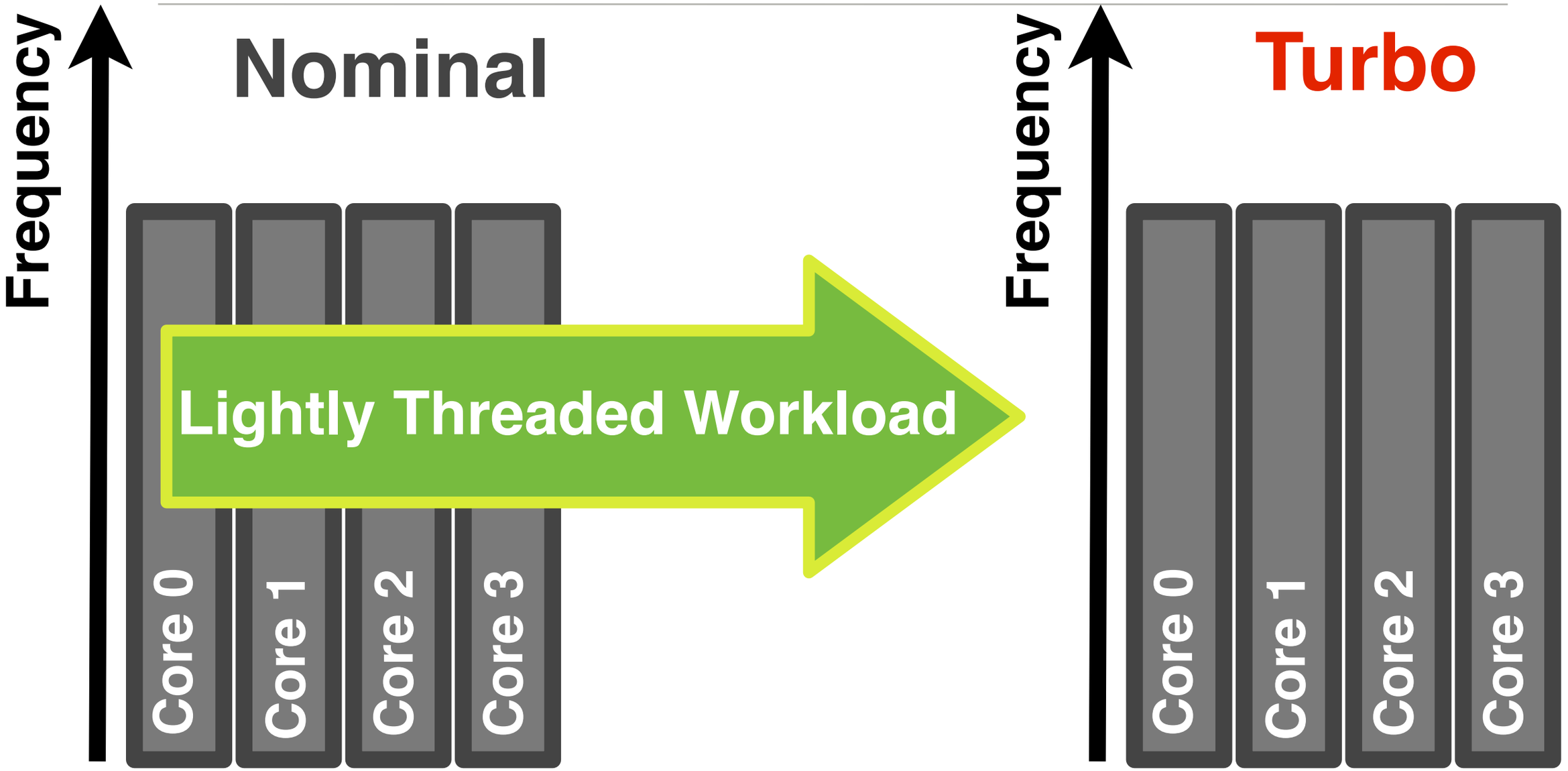
Example V/f Boosting: Intel Turbo Boost



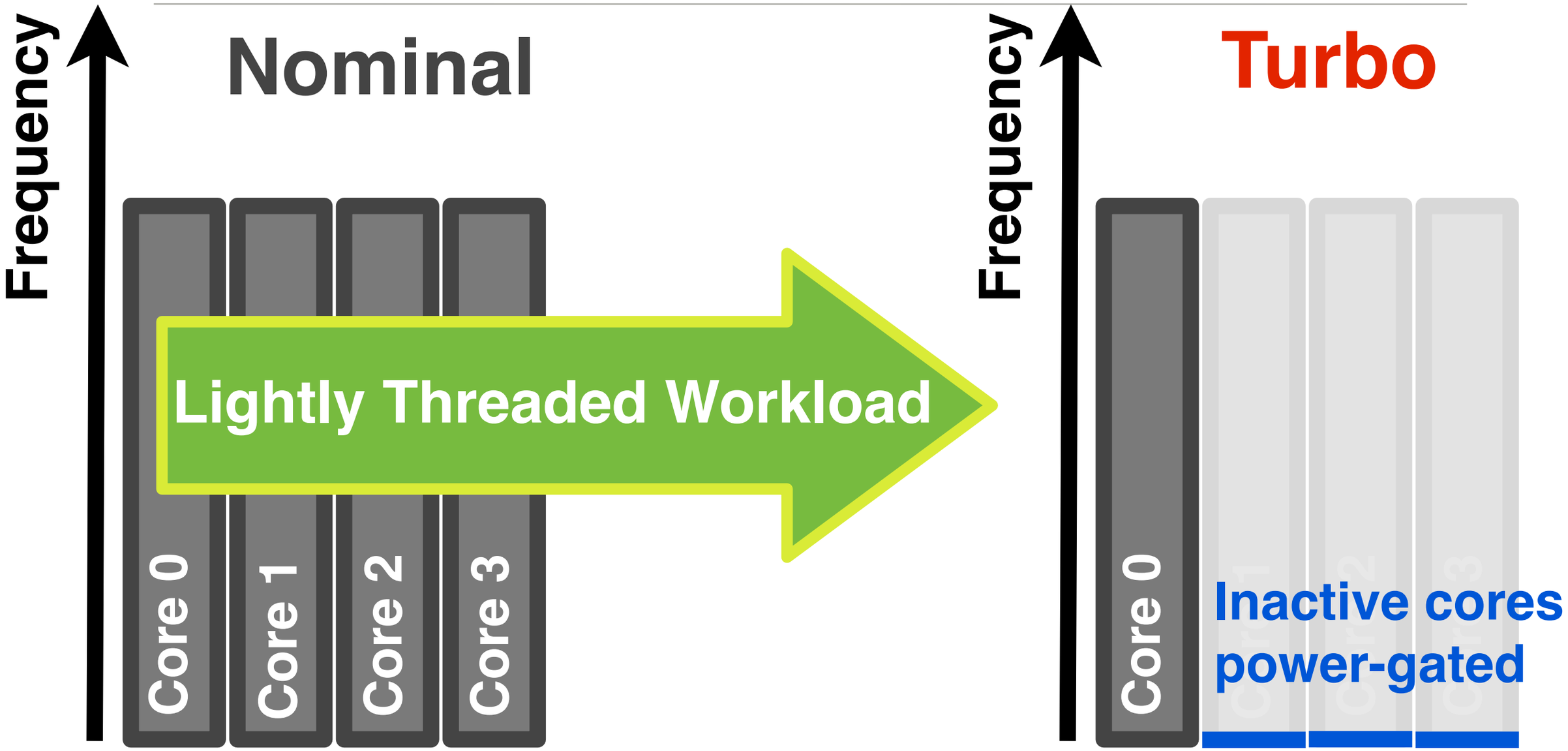
Example V/f Boosting: Intel Turbo Boost



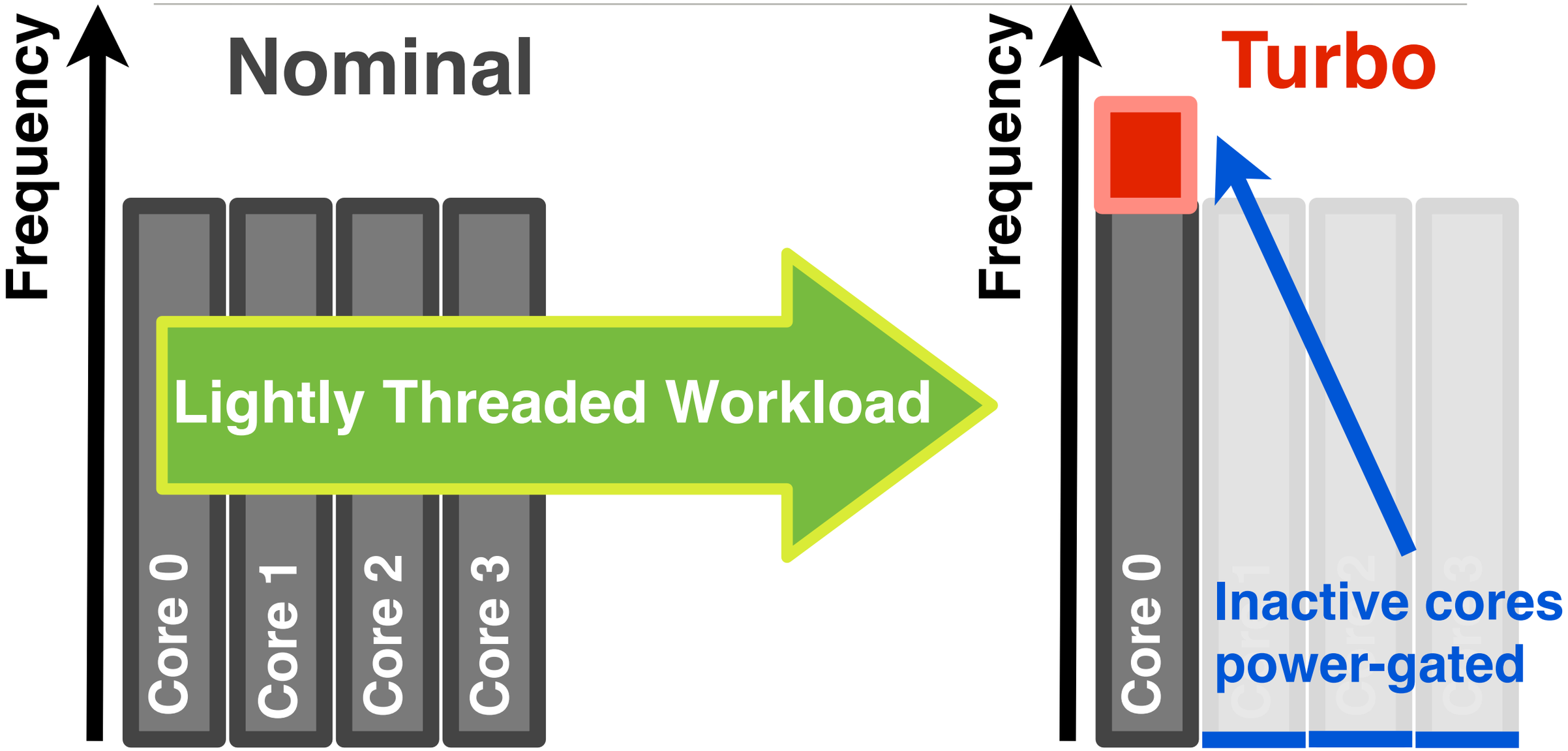
Example V/f Boosting: Intel Turbo Boost



Example V/f Boosting: Intel Turbo Boost



Example V/f Boosting: Intel Turbo Boost



Composing the Techniques

Composing the Techniques

The techniques are orthogonal

Composing the Techniques

The techniques are orthogonal

- Suppose as much P/T headroom as necessary

Composing the Techniques

The techniques are orthogonal

- Suppose as much P/T headroom as necessary
- Paceline: Performance gain constrained by P_{EMAX}

Composing the Techniques

The techniques are orthogonal

- Suppose as much P/T headroom as necessary
- Paceline: Performance gain constrained by P_{EMAX}
- V/f Boosting: Performance gain constrained by V_{MAX}

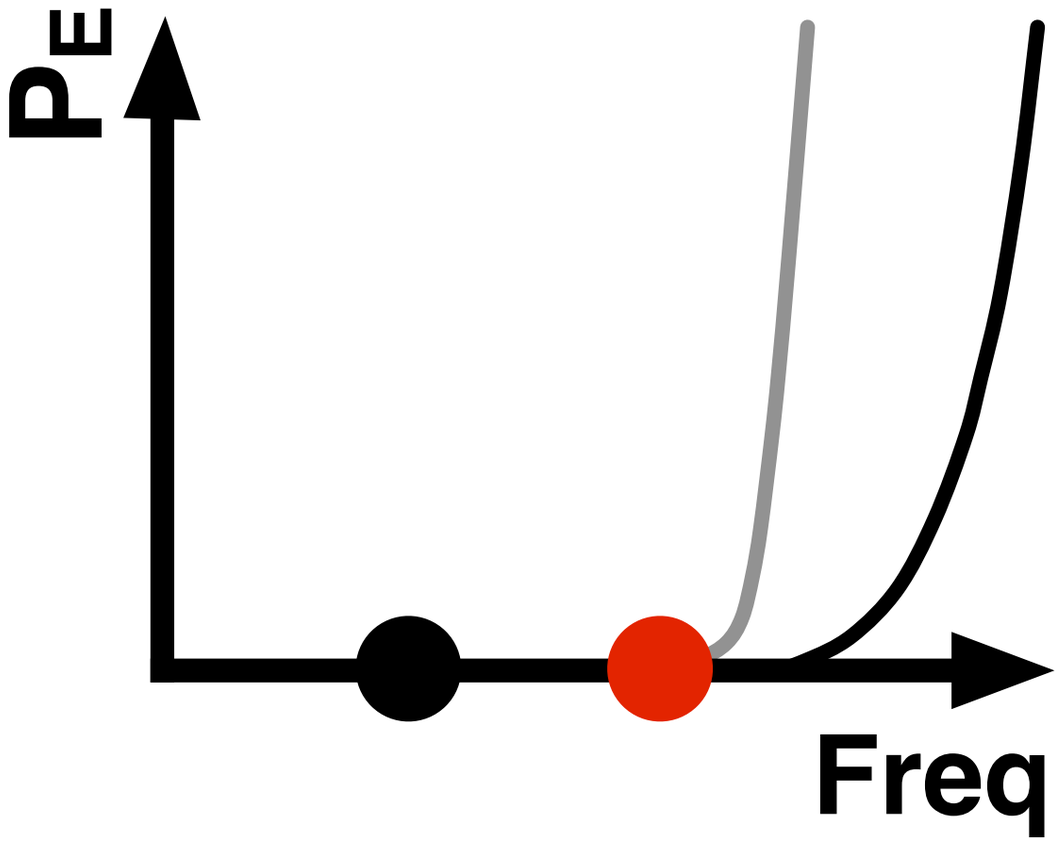
V/f Boosting + Paceline

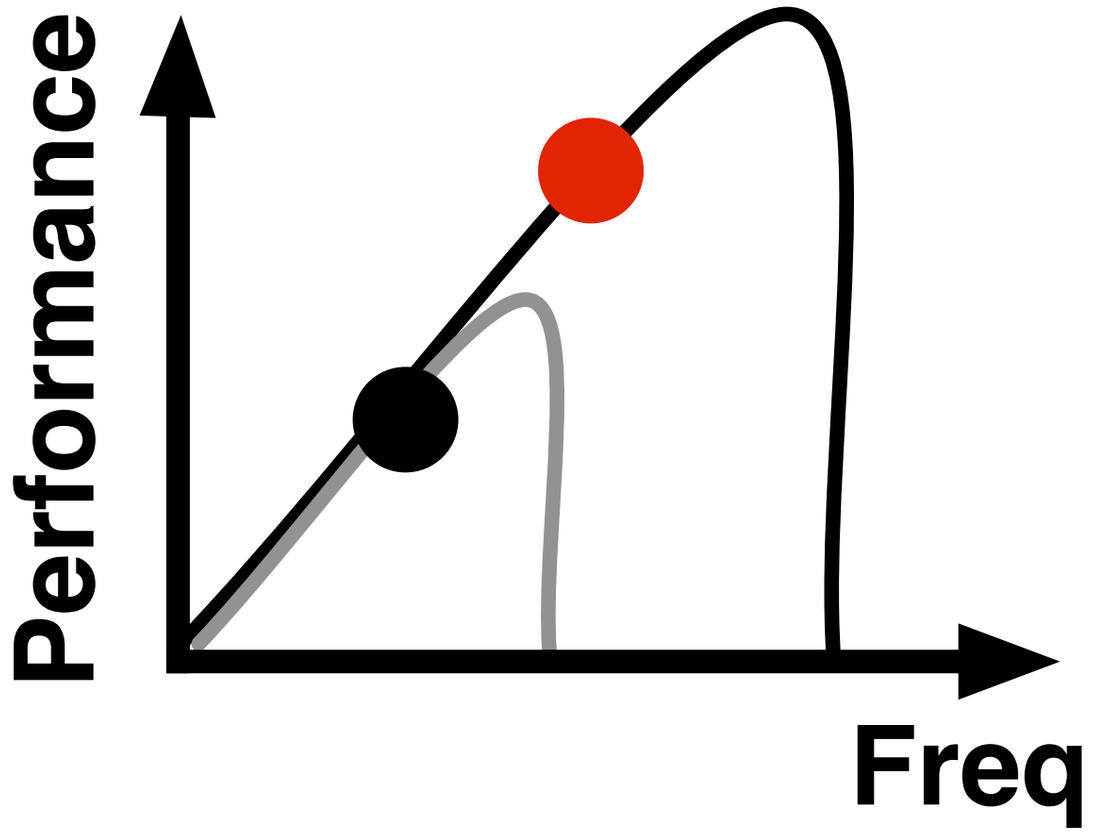
V/f Boosting + Paceline

Boost core frequency beyond nominal by **increasing** V; tolerating occasional timing errors

V/f Boosting + Paceline

Boost core frequency beyond nominal by **increasing** V; tolerating occasional timing errors



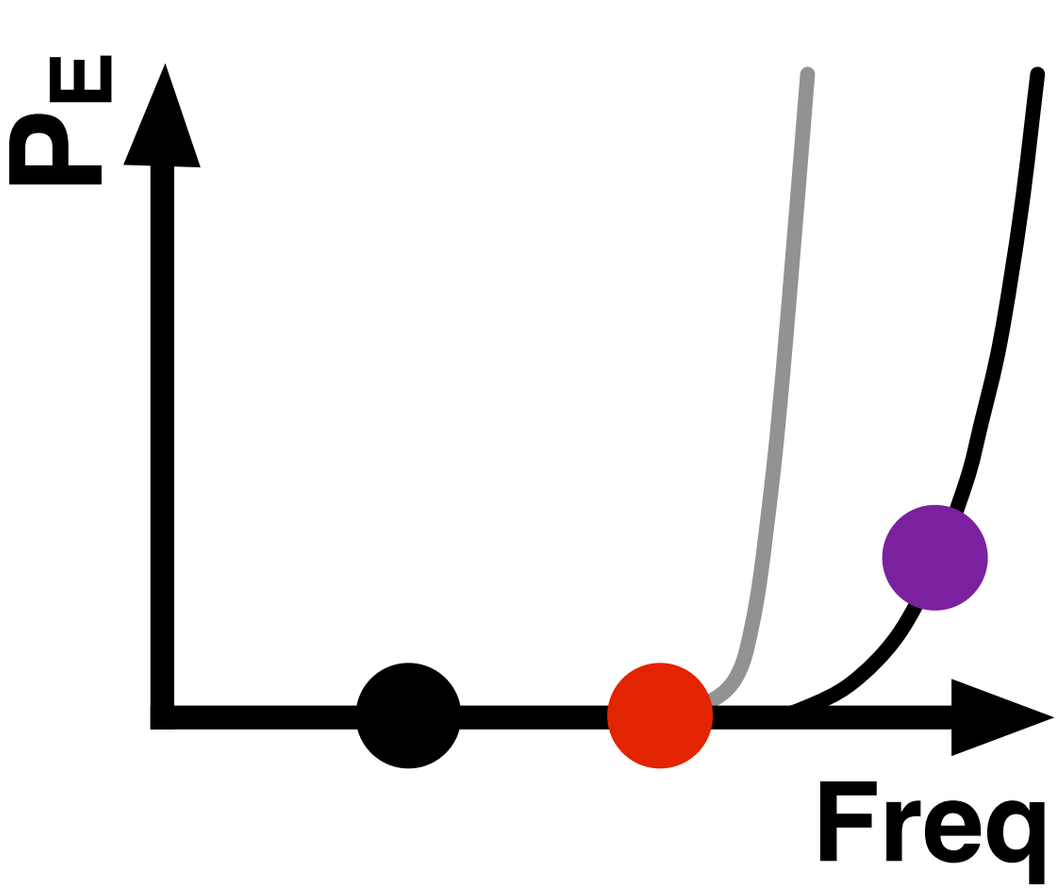


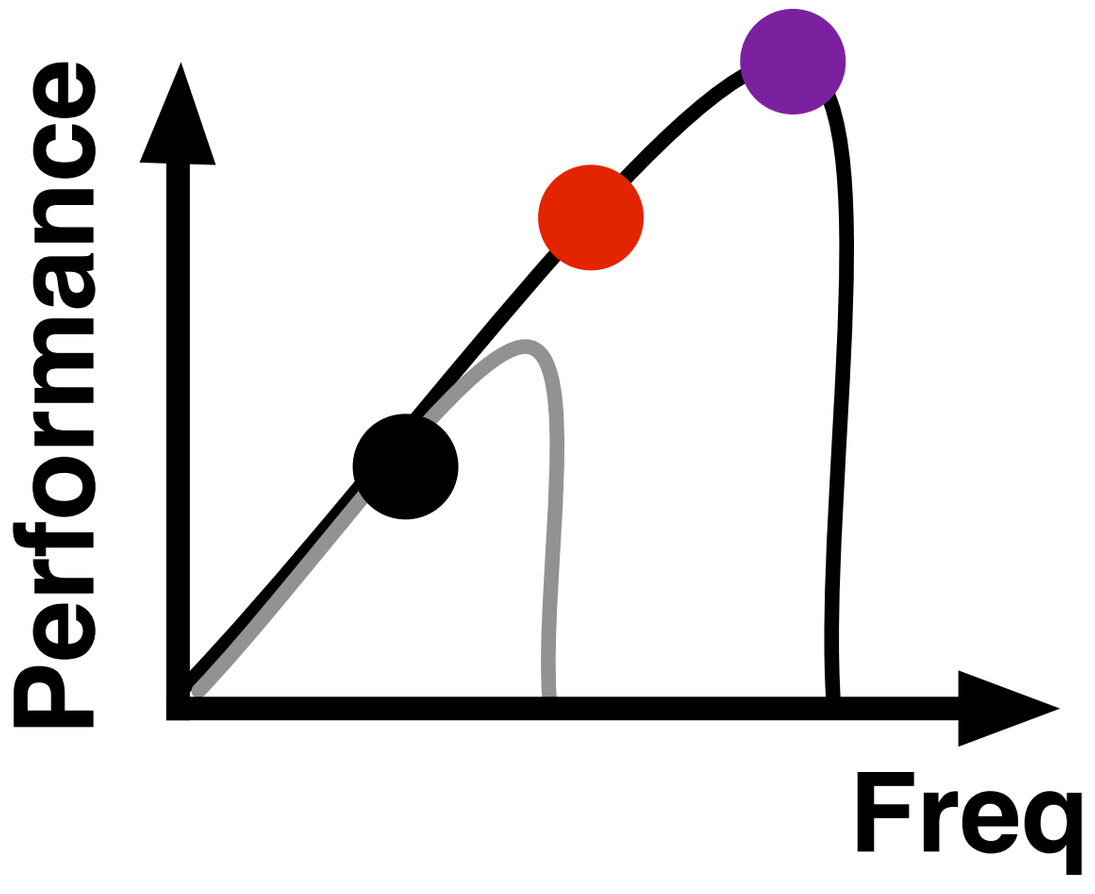
Rated

V/f Boosting

V/f Boosting + Paceline

Boost core frequency beyond nominal by **increasing** V; tolerating occasional timing errors





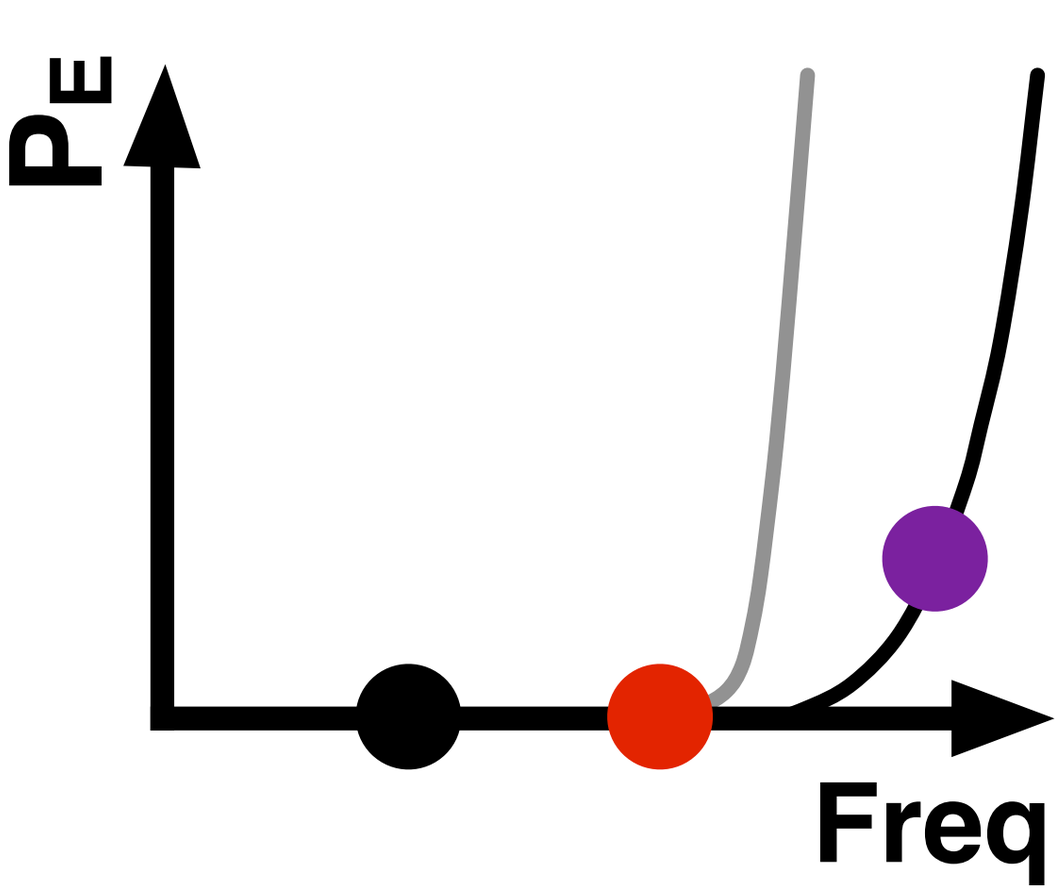
Rated

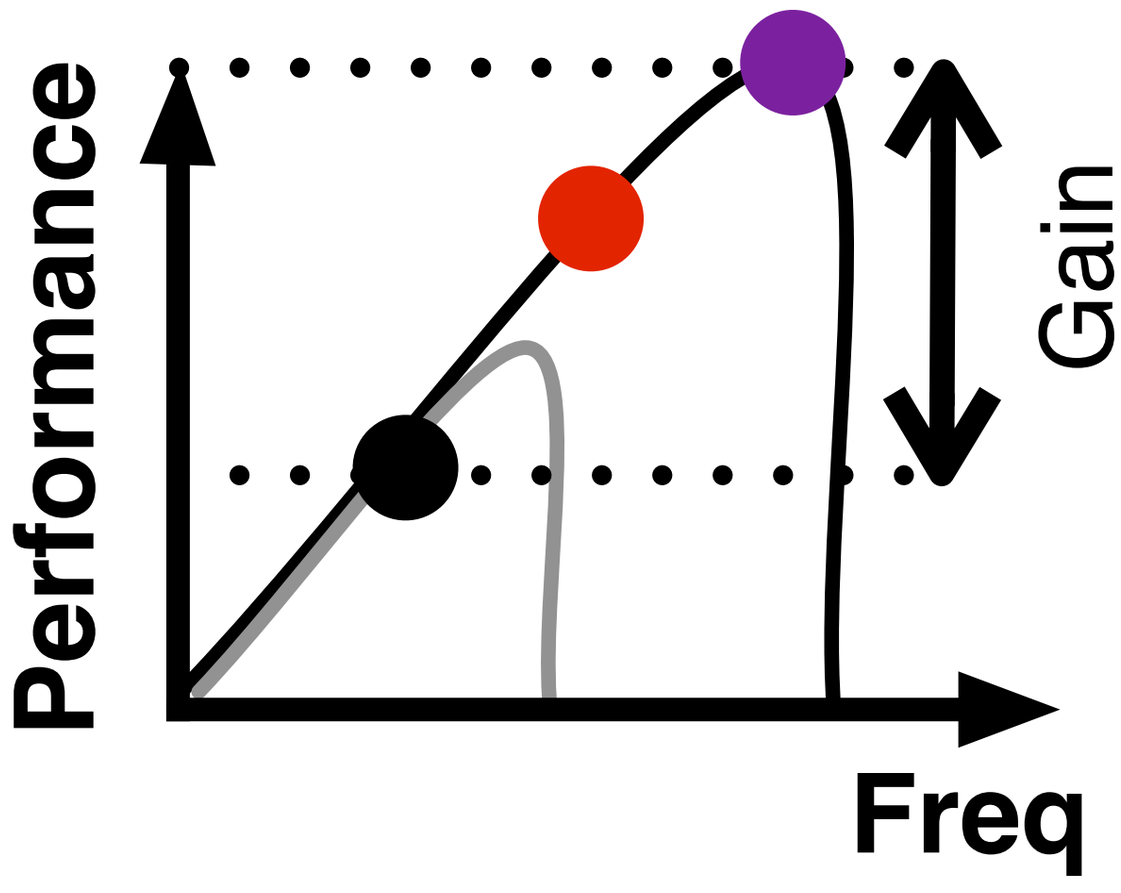
V/f Boosting

V/f Boosting + Paceline

V/f Boosting + Paceline

Boost core frequency beyond nominal by **increasing** V; tolerating occasional timing errors





Rated

V/f Boosting

V/f Boosting + Paceline

Composing the Techniques

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		Unlikely

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		Unlikely
High to Moderate	✓			✓	✓		

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		Unlikely
High to Moderate	✓			✓	✓		

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		Unlikely
High to Moderate	✓			✓	✓		

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		Unlikely
High to Moderate	✓			✓	✓		

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		Unlikely
High to Moderate	✓			✓	✓		**Likely**

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		Unlikely
High to Moderate		✓		✓	✓		Likely
Low		✓		✓		✓	

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		Unlikely
High to Moderate		✓		✓	✓		Likely
Low		✓		✓		✓	

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		Unlikely
High to Moderate		✓		✓	✓		Likely
Low		✓		✓		✓	

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		Unlikely
High to Moderate		✓		✓	✓		Likely
Low		✓		✓		✓	

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		Unlikely
High to Moderate		✓		✓	✓		Likely
Low		✓		✓		✓	**Definitely**

Composing the Techniques

Multicore Loading Condition	Bounding Constraints						Gain from combining ?
	V/f Boost		Paceline		Compos.		
	T/P	V	T/P	P_E	T/P	V/P_E	
Very High	✓		✓		✓		Unlikely
High to Moderate		✓		✓	✓		Likely
Low		✓		✓		✓	**Definitely**

Additional techniques can be applied

LeadOut Operation

LeadOut Operation

- At any given time, CMP executes a mix of speed-critical and throughput-oriented threads

LeadOut Operation

- At any given time, CMP executes a mix of speed-critical and throughput-oriented threads
- Run throughput threads unoptimized

LeadOut Operation

- At any given time, CMP executes a mix of speed-critical and throughput-oriented threads
- Run throughput threads unoptimized
- Optimization Problem

LeadOut Operation

- At any given time, CMP executes a mix of speed-critical and throughput-oriented threads
- Run throughput threads unoptimized
- Optimization Problem

1. Which technique to use for a speed-critical thread?

LeadOut Operation

- At any given time, CMP executes a mix of speed-critical and throughput-oriented threads
- Run throughput threads unoptimized
- Optimization Problem

1. Which technique to use for a speed-critical thread?
2. How to optimally set V/f for chosen technique?

LeadOut Evaluation

LeadOut Evaluation

- Simulated 32nm CMP with 16 cores

LeadOut Evaluation

- Simulated 32nm CMP with 16 cores
- Detailed modeling of leakage, temperature, variation

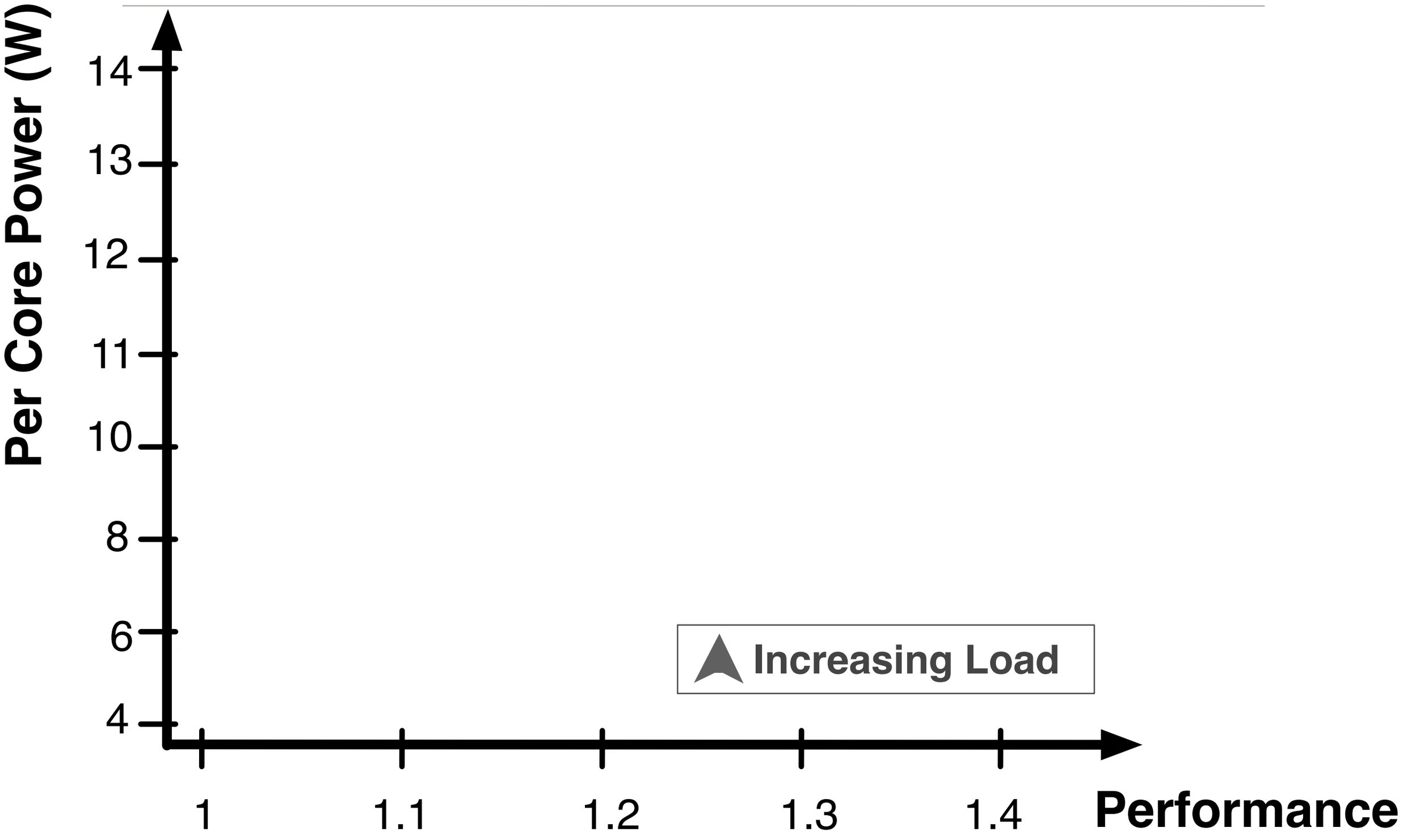
LeadOut Evaluation

- Simulated 32nm CMP with 16 cores
- Detailed modeling of leakage, temperature, variation
- Applications: SPECint2000 benchmarks

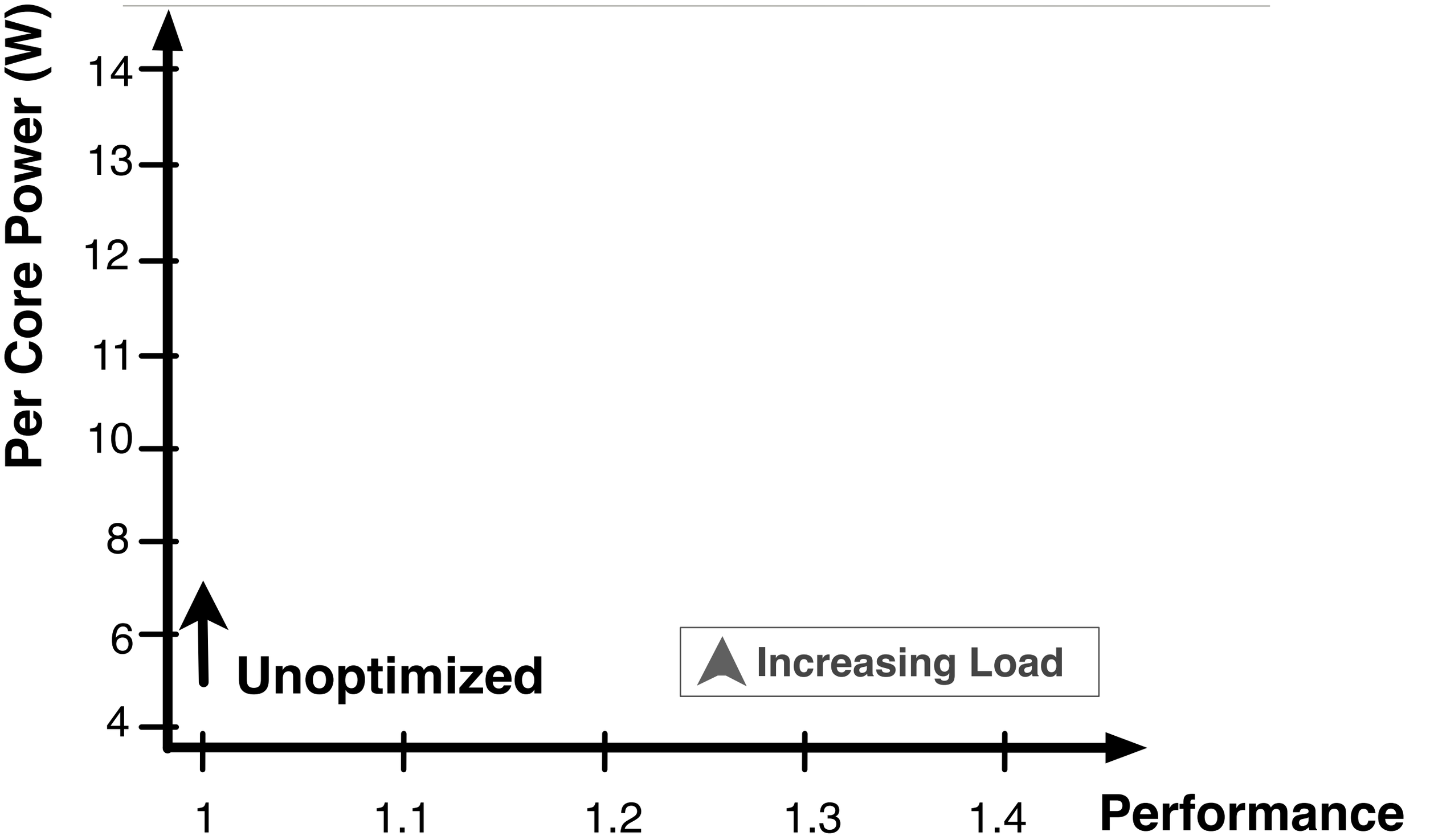
LeadOut Evaluation

- Simulated 32nm CMP with 16 cores
- Detailed modeling of leakage, temperature, variation
- Applications: SPECint2000 benchmarks
- 50 Monte Carlo die samples

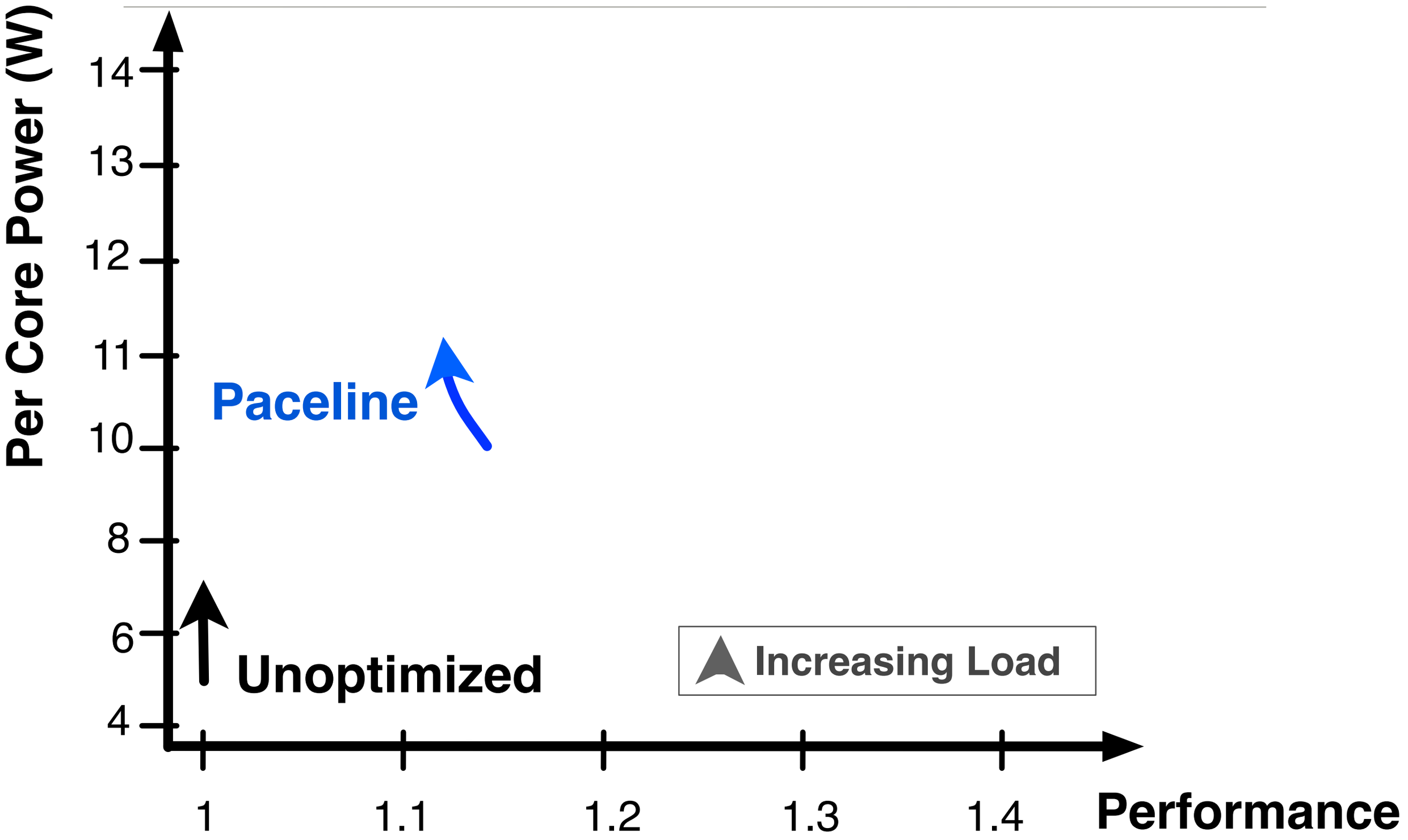
Power-Performance Tradeoff



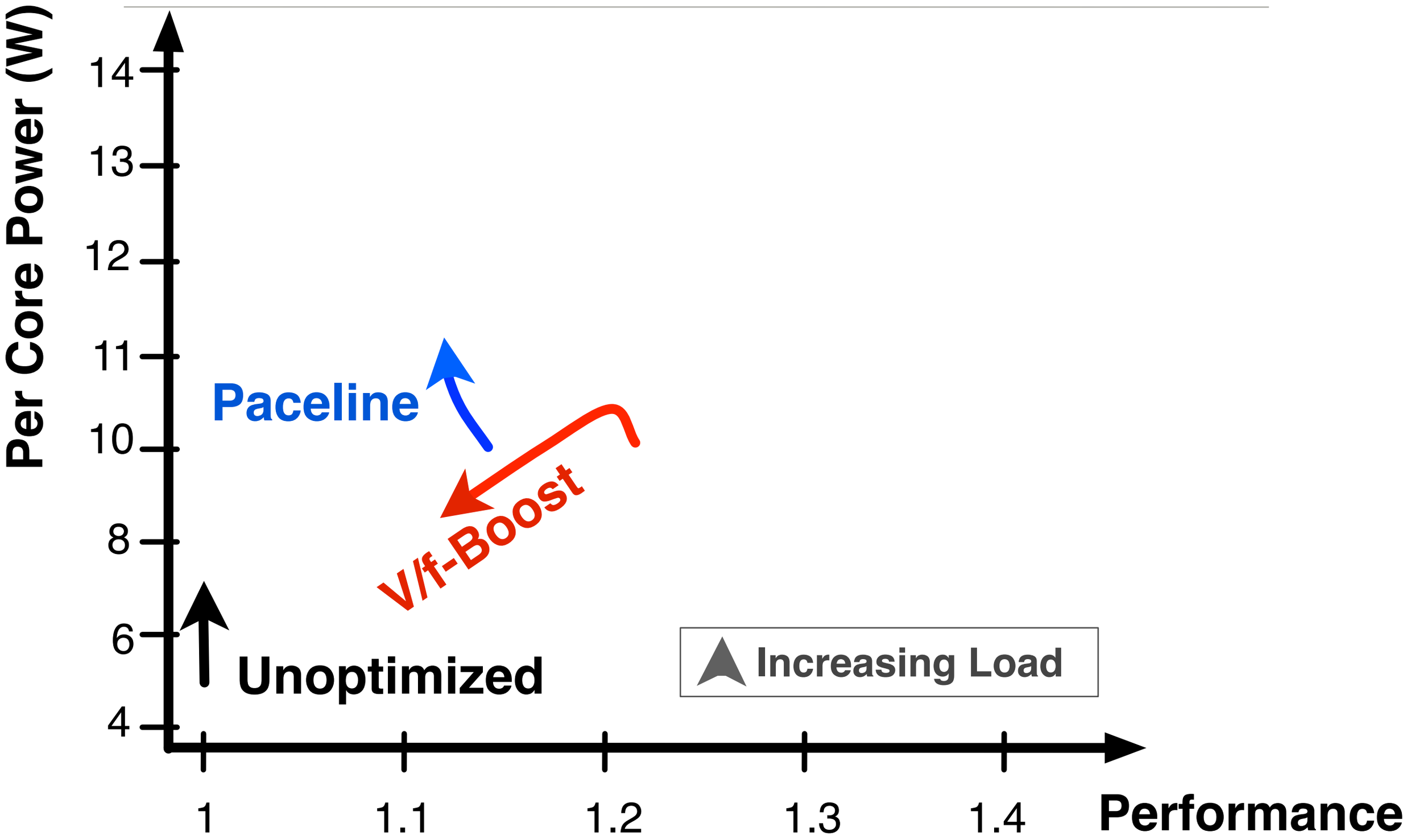
Power-Performance Tradeoff



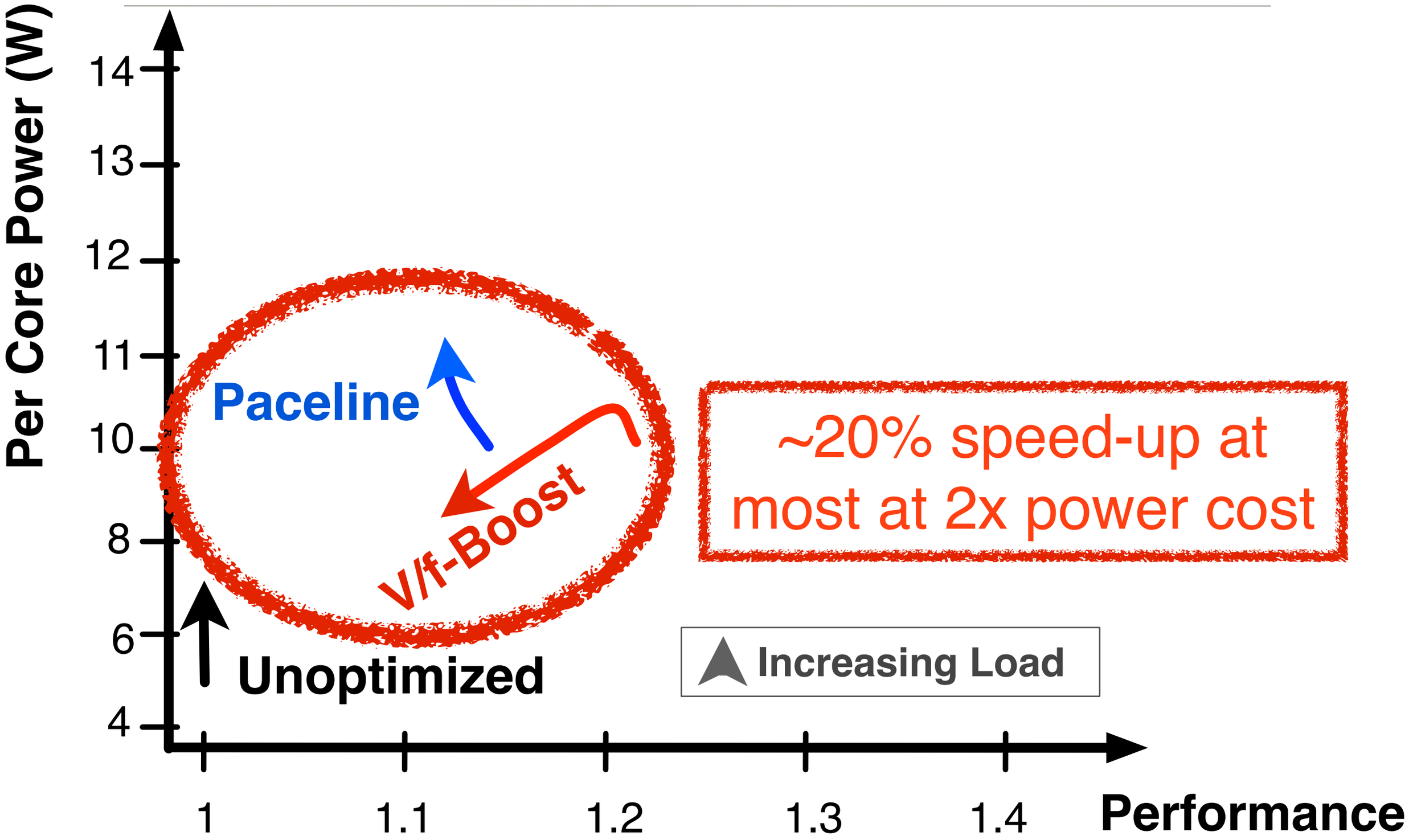
Power-Performance Tradeoff



Power-Performance Tradeoff

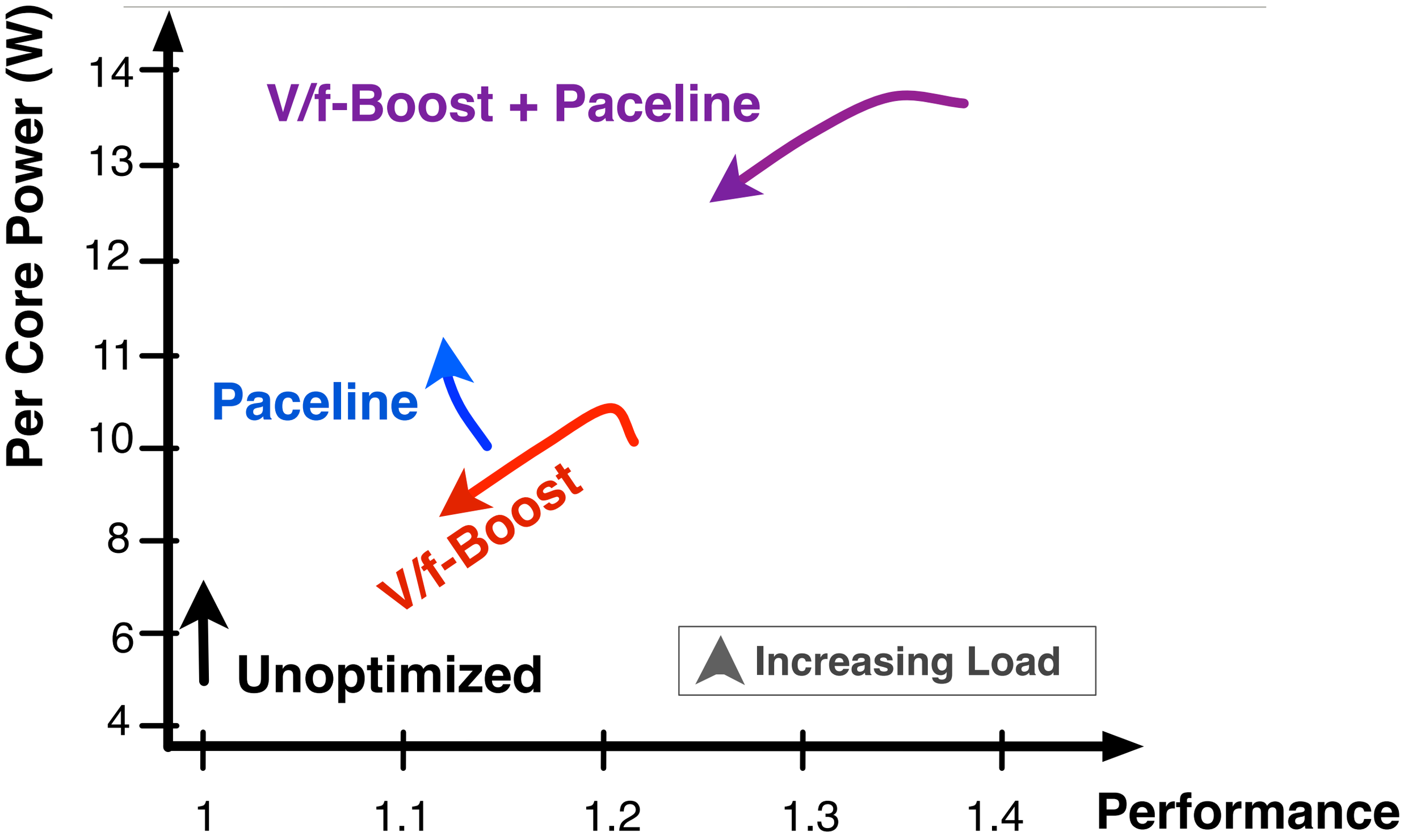


Power-Performance Tradeoff

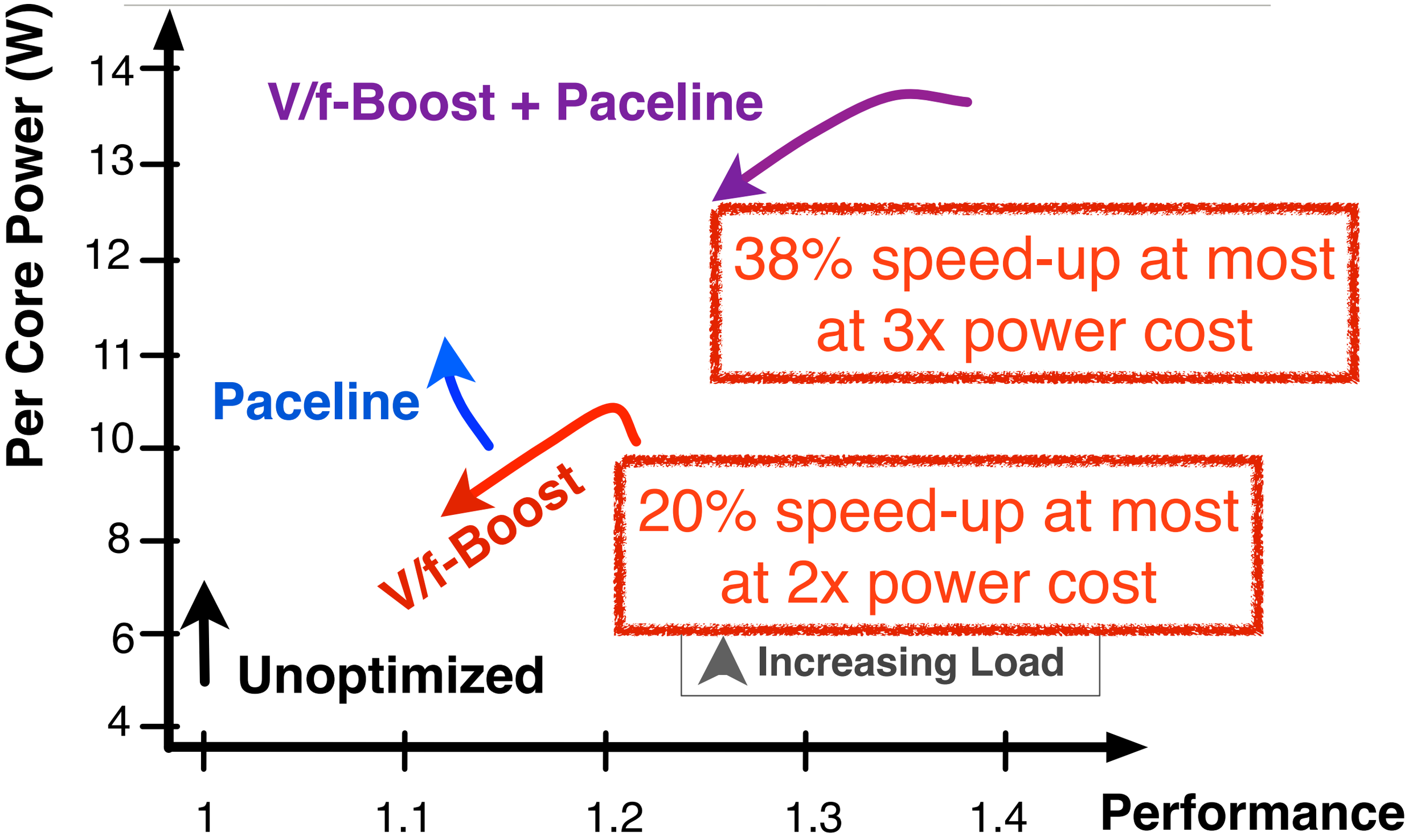


~20% speed-up at most at 2x power cost

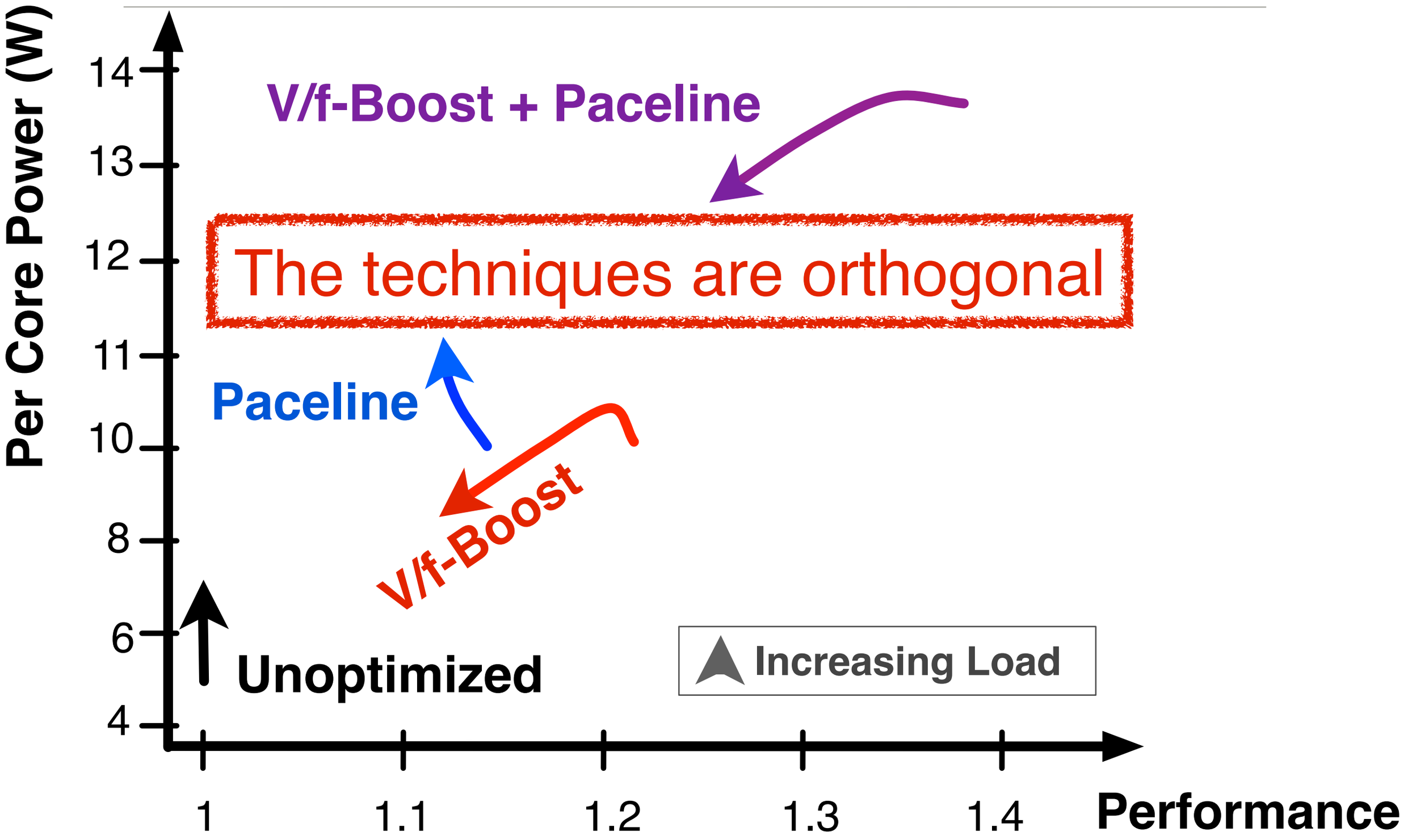
Power-Performance Tradeoff



Power-Performance Tradeoff



Power-Performance Tradeoff



Also in the Paper

Also in the Paper

- Detailed analysis for different load conditions

Also in the Paper

- Detailed analysis for different load conditions
  - Including V/f Boosting with activity migration

Also in the Paper

- Detailed analysis for different load conditions
  - Including V/f Boosting with activity migration
- Sensitivity analysis

Also in the Paper

- Detailed analysis for different load conditions
  - Including V/f Boosting with activity migration
- Sensitivity analysis
  - Thermal design points, power grid designs, guardband

Also in the Paper

- Detailed analysis for different load conditions
  - Including V/f Boosting with activity migration
- Sensitivity analysis
  - Thermal design points, power grid designs, guardband
- A hierarchical controller design to dynamically set

Also in the Paper

- Detailed analysis for different load conditions
  - Including V/f Boosting with activity migration
- Sensitivity analysis
  - Thermal design points, power grid designs, guardband
- A hierarchical controller design to dynamically set
  - Technique to apply

Also in the Paper

- Detailed analysis for different load conditions
  - Including V/f Boosting with activity migration
- Sensitivity analysis
  - Thermal design points, power grid designs, guardband
- A hierarchical controller design to dynamically set
  - Technique to apply
  - Per core V/f assignment for the chosen technique

Conclusion

Conclusion

- Two low overhead techniques for sequential acceleration

Conclusion

- Two low overhead techniques for sequential acceleration
  - V/f Boosting and Timing Speculation

Conclusion

- Two low overhead techniques for sequential acceleration
  - V/f Boosting and Timing Speculation
  - Individual application of any technique suboptimal

Conclusion

- Two low overhead techniques for sequential acceleration
  - V/f Boosting and Timing Speculation
  - Individual application of any technique suboptimal
  - Techniques are complementary

Conclusion

- Two low overhead techniques for sequential acceleration
  - V/f Boosting and Timing Speculation
  - Individual application of any technique suboptimal
  - Techniques are complementary
- LeadOut: A highly-configurable CMP

Conclusion

- Two low overhead techniques for sequential acceleration
  - V/f Boosting and Timing Speculation
  - Individual application of any technique suboptimal
  - Techniques are complementary
- LeadOut: A highly-configurable CMP
  - Combining V/f Boosting and TS synergistically

Conclusion

- Two low overhead techniques for sequential acceleration
  - V/f Boosting and Timing Speculation
  - Individual application of any technique suboptimal
  - Techniques are complementary
- LeadOut: A highly-configurable CMP
  - Combining V/f Boosting and TS synergistically
  - 34% thread speedup at 220% power increase

LeadOut: Composing Low-Overhead Techniques for Single-Thread Performance

Brian Greskamp, **Ulya Karpuzcu**, Josep Torrellas

http://iacoma.cs.uiuc.edu/

University
of Illinois

i-acoma group