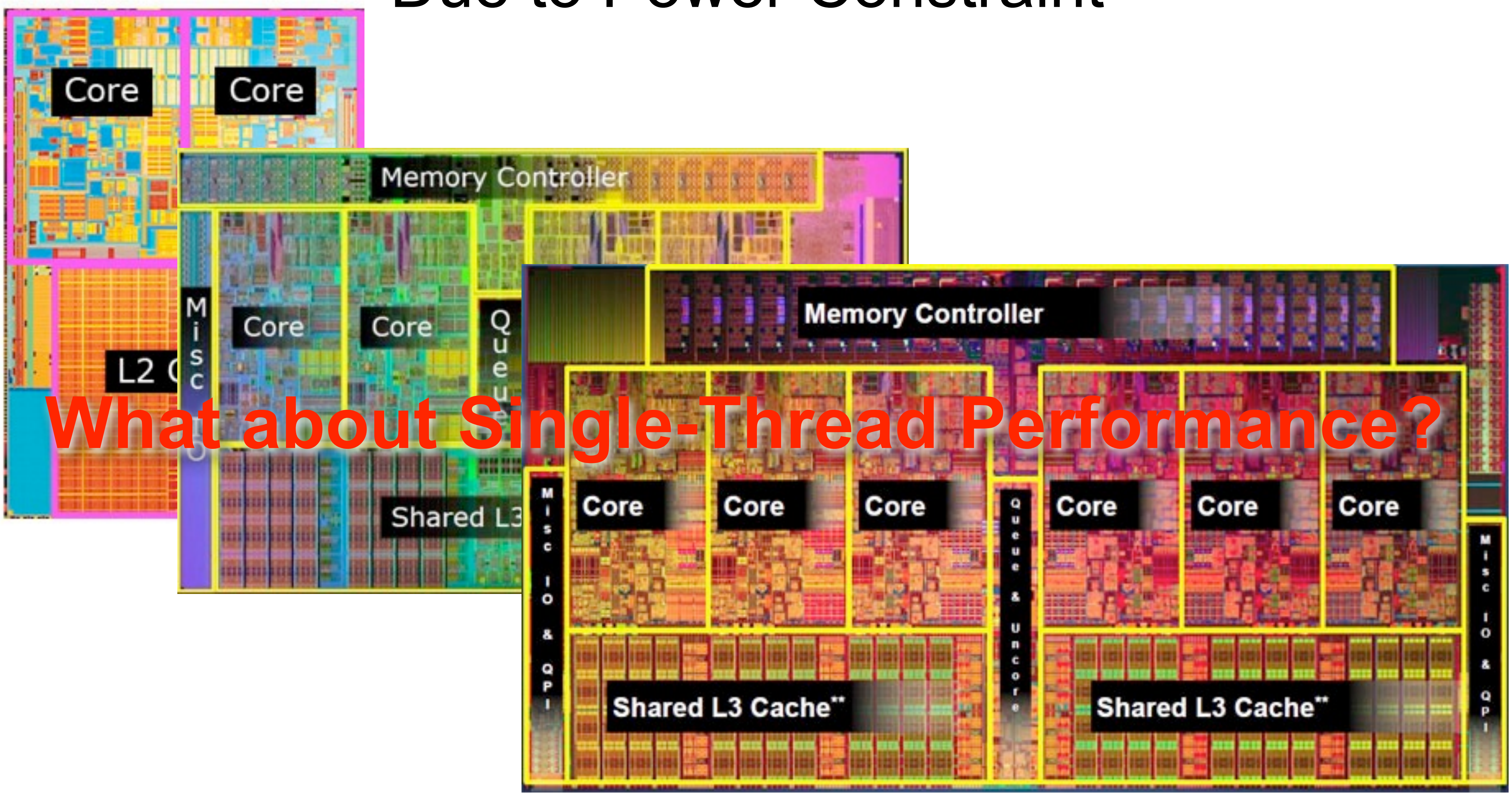
Analytical Latency-Throughput Model of Future Power Constrained Multicore Processors

Amanda Chih-Ning Tseng and David Brooks
Harvard University

WEED 2012 June 9th, 2012

For Today's Multicore Processors, the Trend Towards Maximizing Throughput Due to Power Constraint



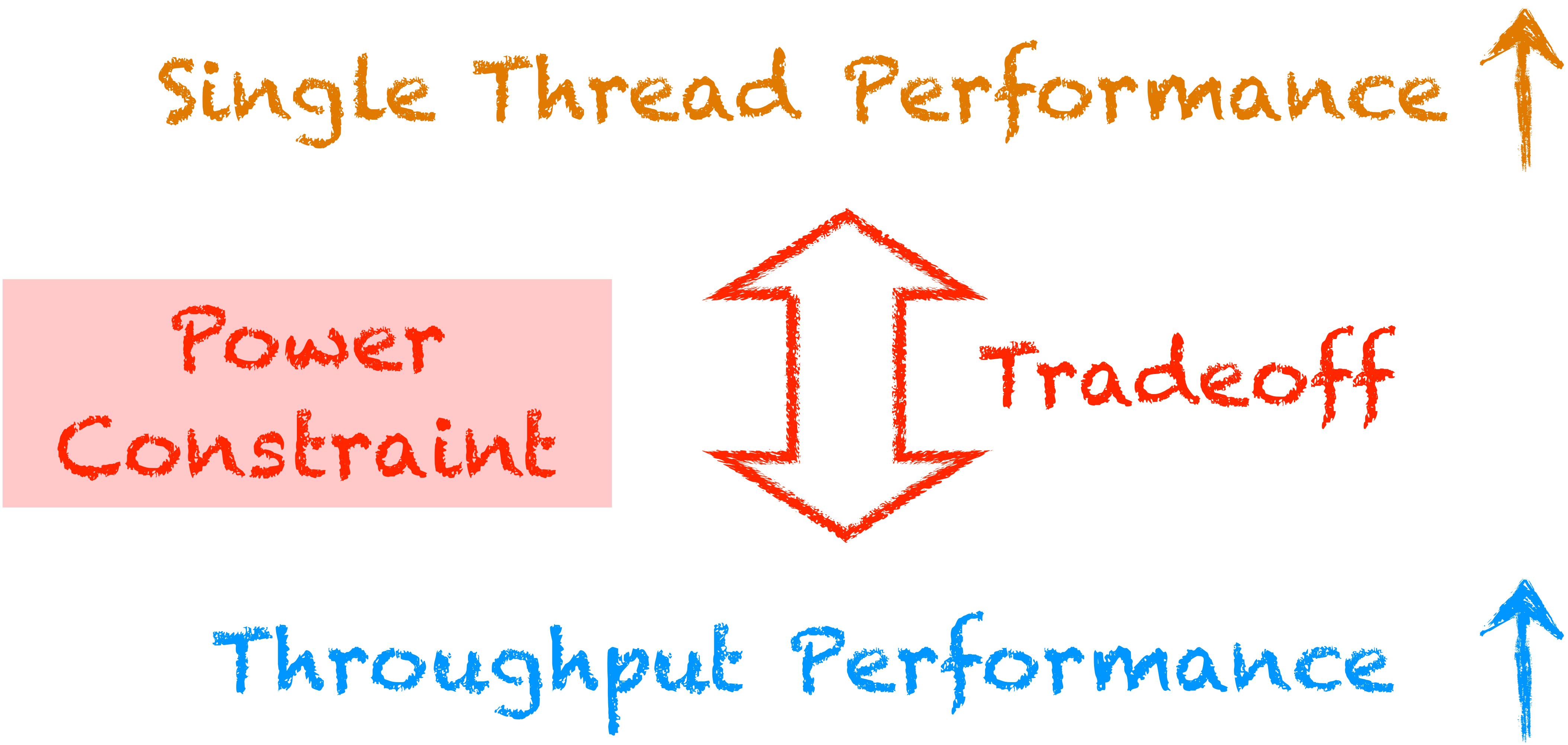
What about Single-Thread Performance?

App Type	Office	Media Playback	Game	Web Browsing	Image	Video
Avg. TLP	1.2	1.5	1.6	2.0	2.1	7.4

* Blake et. al., *Evolution of Thread-Level Parallelism in Desktop Applications, ISCA'10*

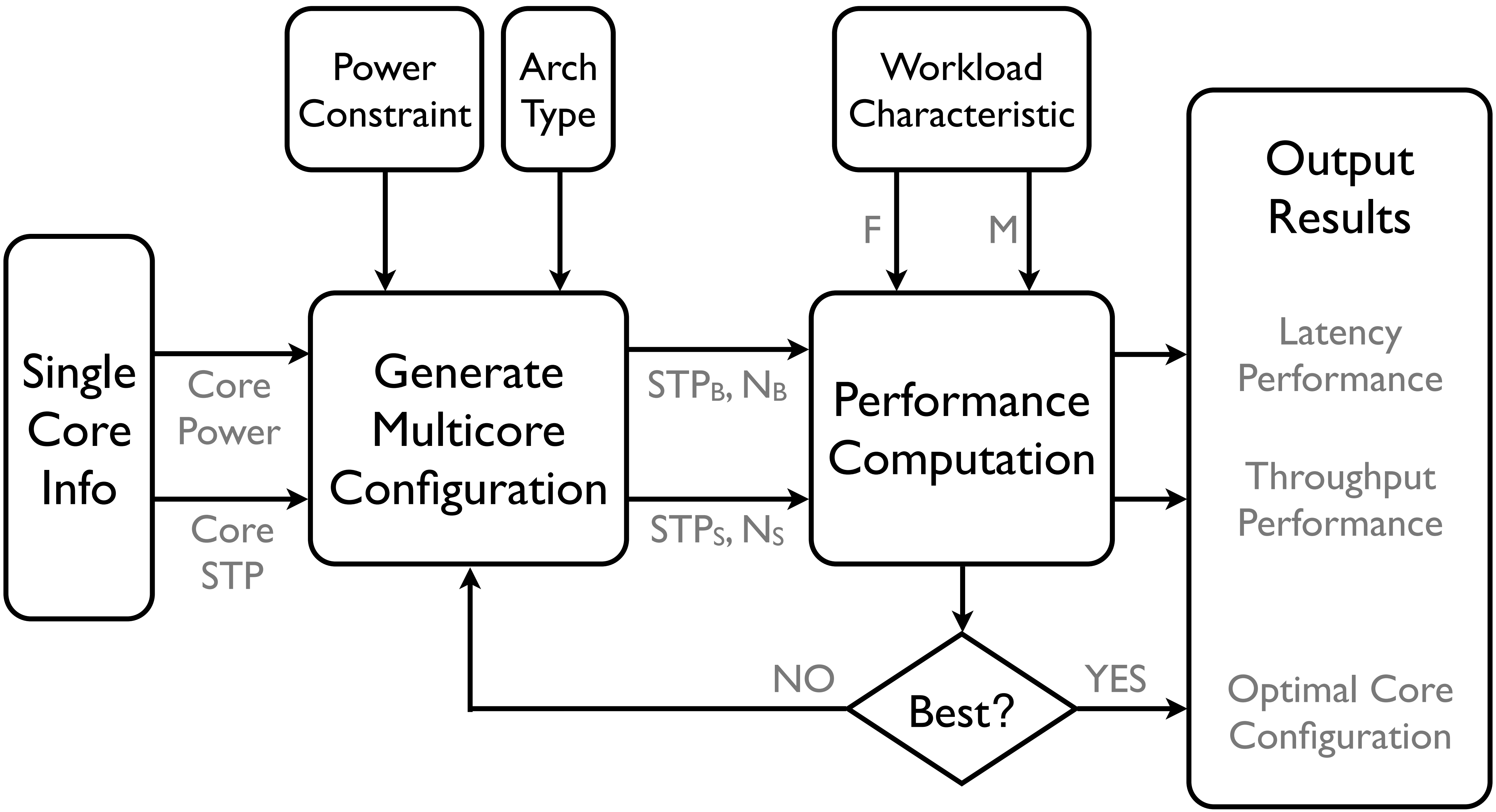
- Most popular desktop and mobile applications benefit less from multicore processors
- Processor performance increased 22%/year for the past decade compared to the historical 52%**

** Hennessy and Patterson, *Computer Architecture, A Quantitative Approach, Ed.5*



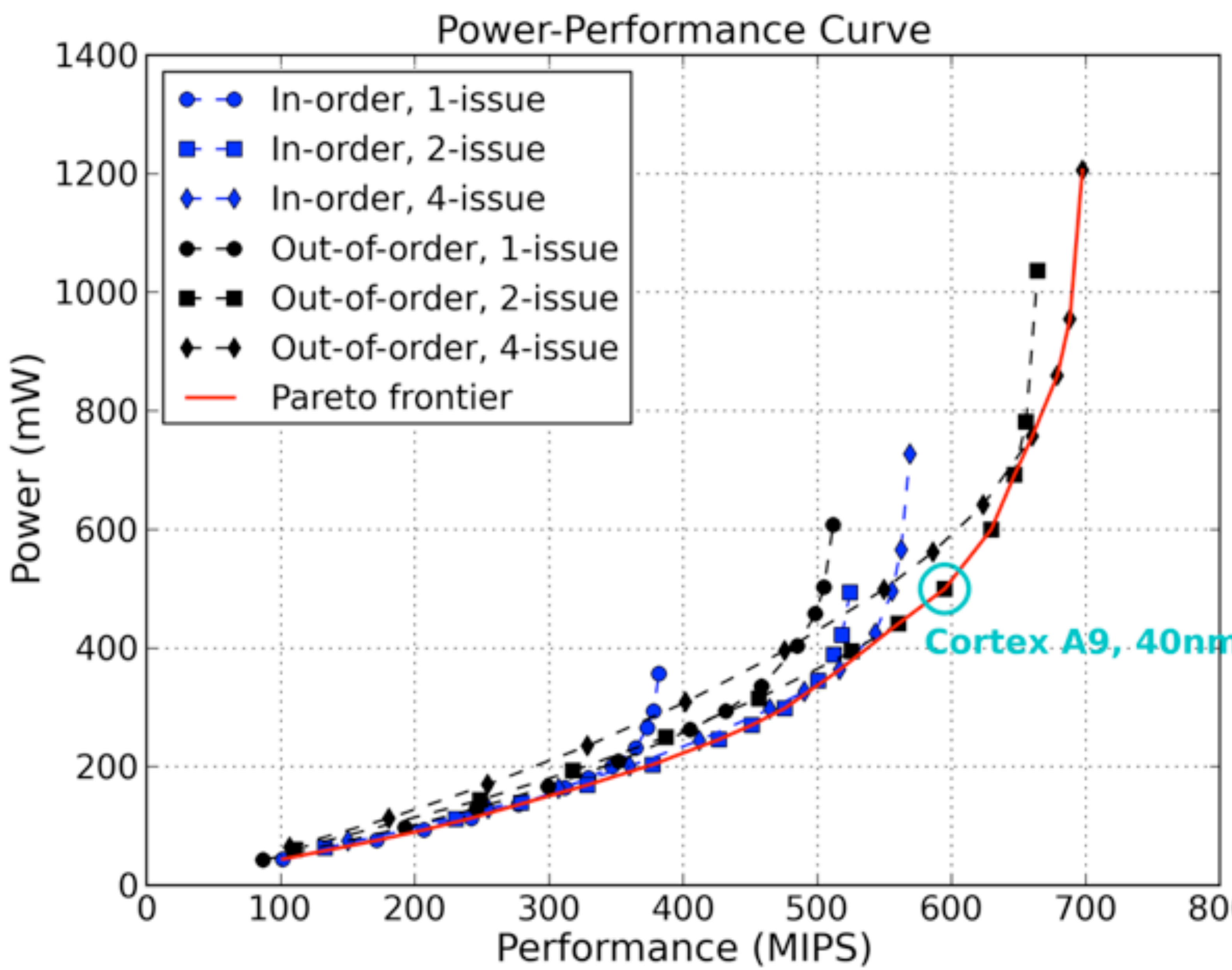
To study the tradeoff, a model that analyzes the performance relation is necessary.

Model Overview



Model Overview

Single Core Info



* Azizi et. al., *Energy-Performance Tradeoffs in Processor Architecture and Circuit Design: A Marginal Cost Analysis, ISCA '10*

Model Overview

Power Constraint

Thermal Design Power (TDP) :
- Maximum power that the cooling system can dissipate
- Decided at design time

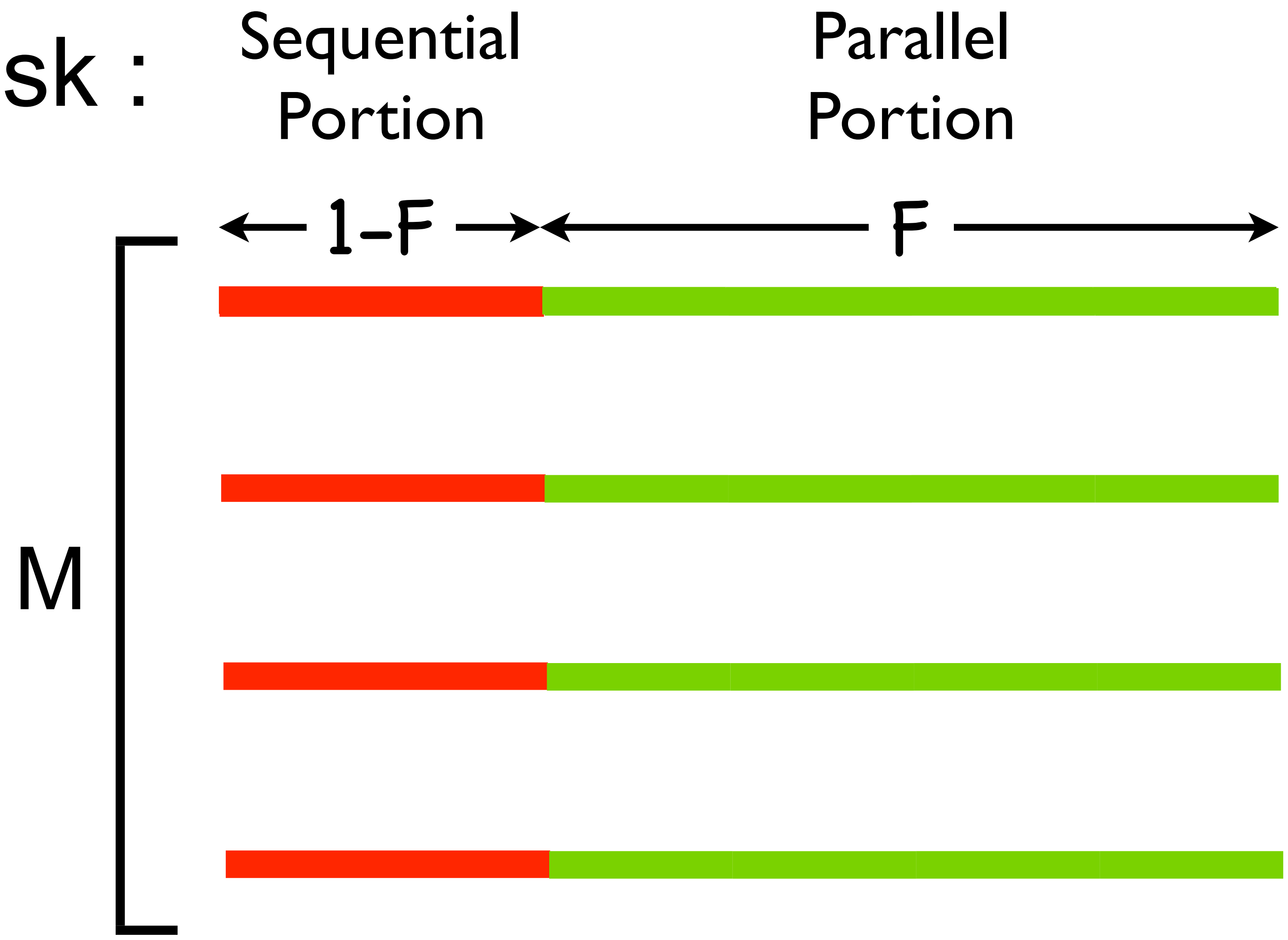
Arch Type

Architecture Type :
- Symmetric or asymmetric multicore
- (Currently implemented)

Model Overview

Workload Characteristic

A Task :



Related Works

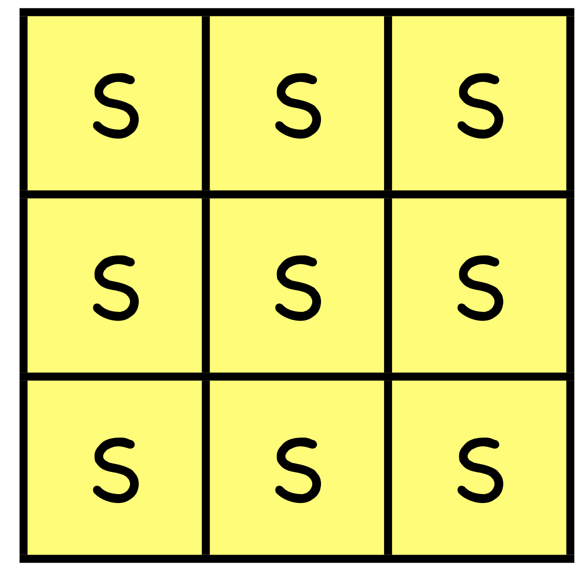
- *“Amdahl's Law in the Multicore Era,”* M. D. Hill and M. R. Marty, Computer, Jul. 2008
  - Augmenting Amdahl’s low to examine theoretical speedups for symmetric, asymmetric, and dynamic multicore chips.
- *“Dark Silicon and the End of Multicore Scaling,”* H. Esmaeilzadeh et. al., ISCA ’11
  - Using technology scaling and performance model to predict speedup and % of dark silicon of future computing systems

Single Task: Symmetric Multicore

$$\text{Amdahl's Law :} \quad \text{Speedup} = \frac{1}{(1-F) + F/N}$$

Power Constraint :

$$N_S = TDP / P_S$$



Performance :

$$STP = STP_S$$

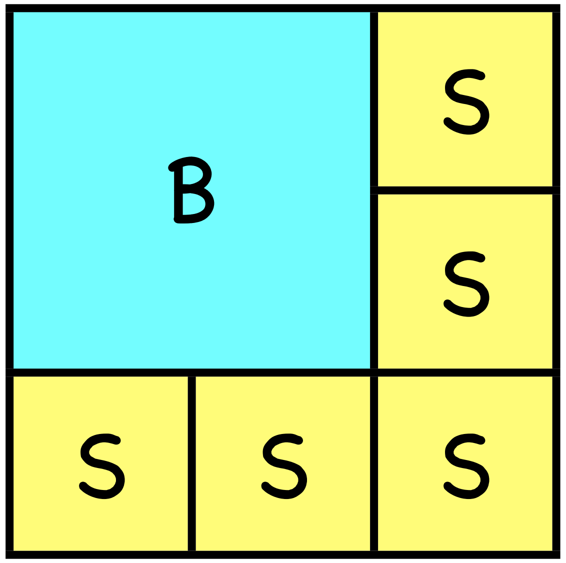
$$THP = STP_S \times \frac{1}{(1-F) + F/N_S} = \frac{1}{\dfrac{(1-F)}{STP_S} + \dfrac{F}{N_S \times STP_S}}$$

Single Task: Asymmetric Multicore

Amdahl's Law : $$\text{Speedup} = \frac{1}{(1-F) + F/N}$$

Power Constraint :

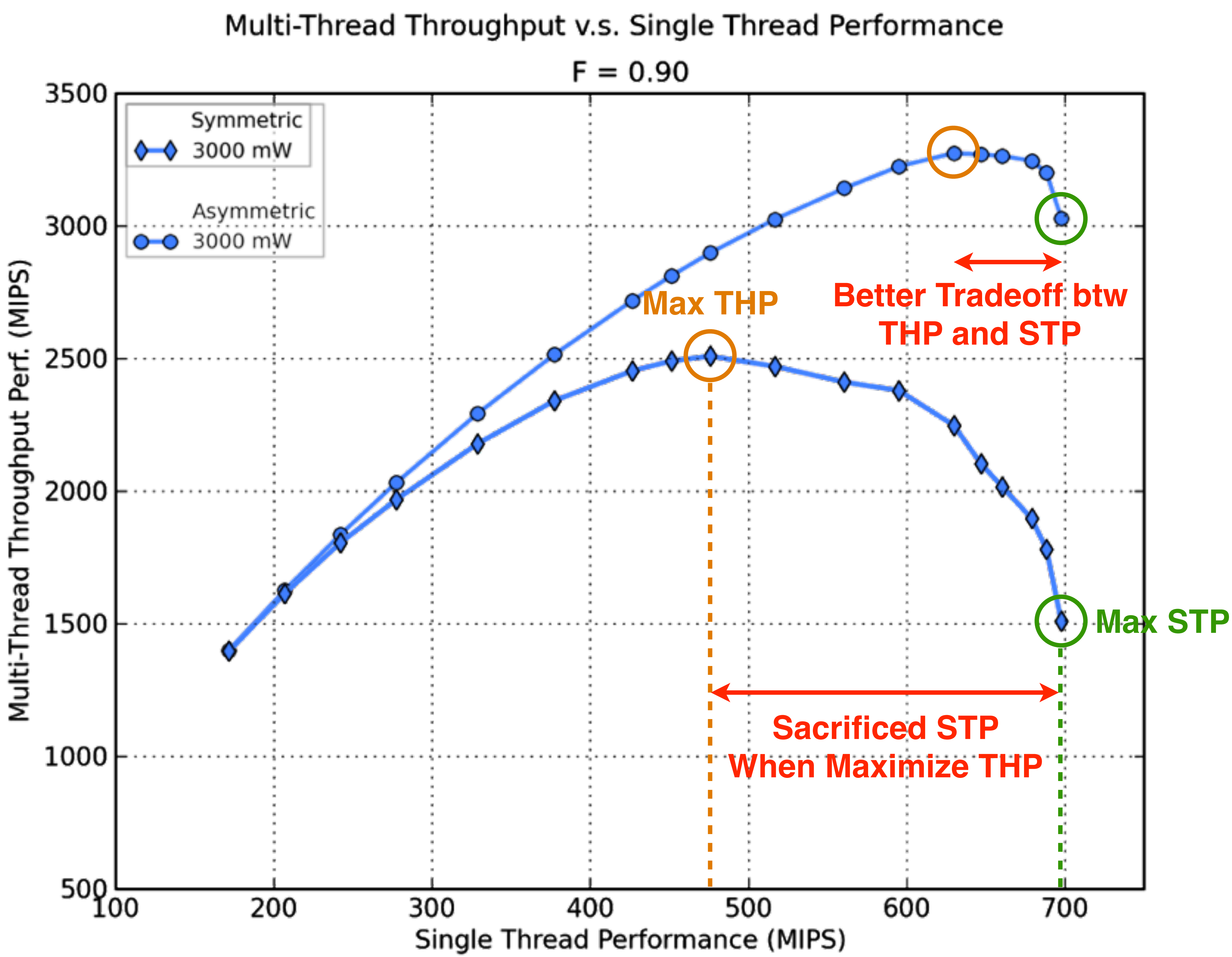
$$N_S = (TDP - P_B) / P_S$$



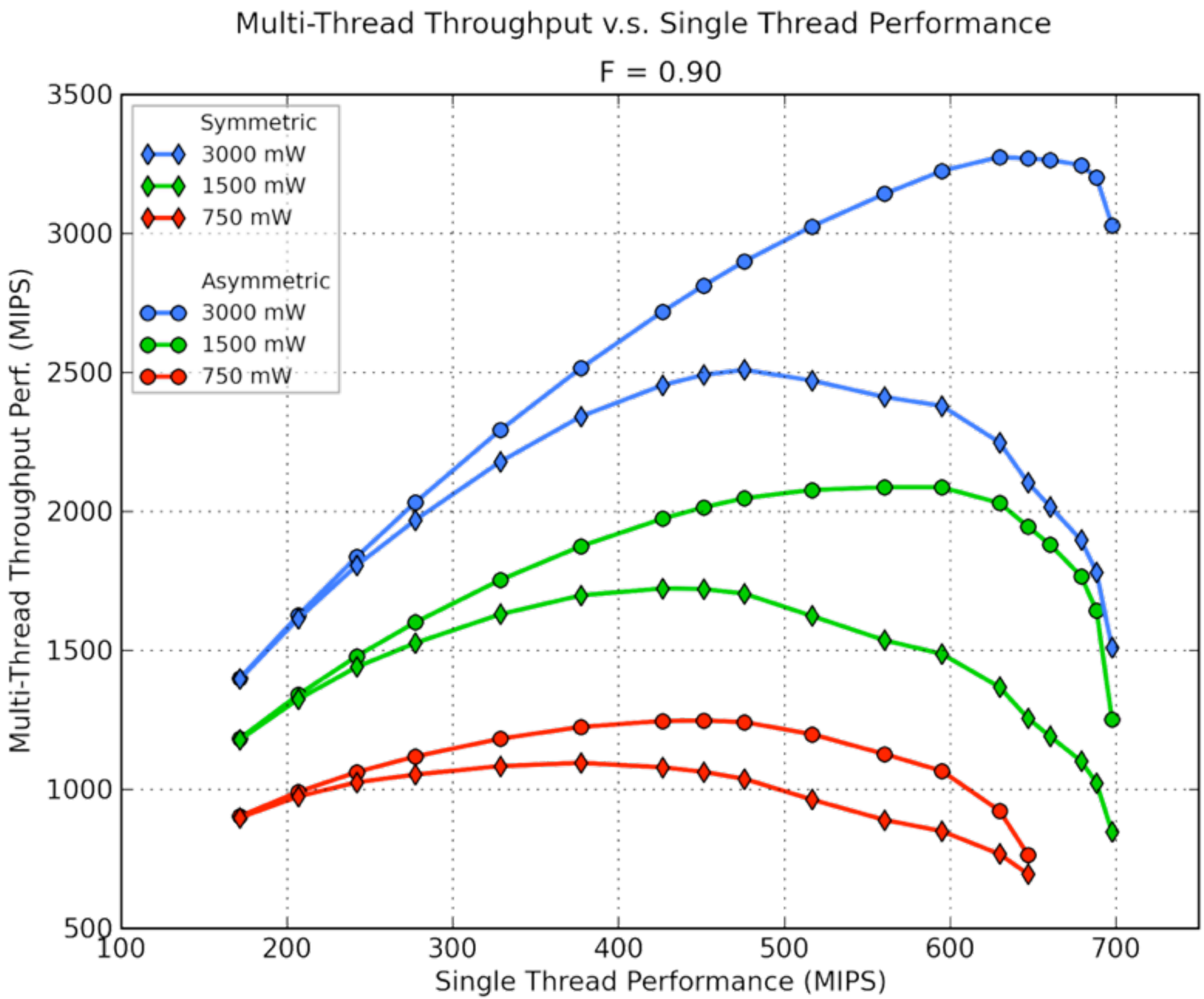
Performance :

$$STP = STP_B$$

$$THP = \frac{1}{\frac{(1-F)}{STP_B} + \frac{F}{STP_B + N_S \times STP_S}}$$



Multi-Thread Throughput v.s. Single Thread Performance
F = 0.90
Symmetric
3000 mW
Asymmetric
3000 mW
Max THP
Better Tradeoff btw
THP and STP
Max STP
Sacrificed STP
When Maximize THP
Multi-Thread Throughput Perf. (MIPS)
Single Thread Performance (MIPS)
3500
3000
2500
2000
1500
1000
500
100
200
300
400
500
600
700



Multi-Thread Throughput v.s. Single Thread Performance
F = 0.90
Symmetric
3000 mW
1500 mW
750 mW
Asymmetric
3000 mW
1500 mW
750 mW
Multi-Thread Throughput Perf. (MIPS)
Single Thread Performance (MIPS)
3500
3000
2500
2000
1500
1000
500
100
200
300
400
500
600
700

Multi-Task Model

- ***Single Task Performance***:
  Throughput performance for single task when using all cores
- ***Multi-Task Throughput Performance***:
  Average time to finish a task when executing multiple tasks
- ***Resource Group***:
  One or more cores exclusively assigned to a particular task

Throughput Generalization

$$THP_{Symm} = \frac{1}{\boxed{\dfrac{(1-F)}{STP_S} + \dfrac{F}{N_S \times STP_S}}}$$

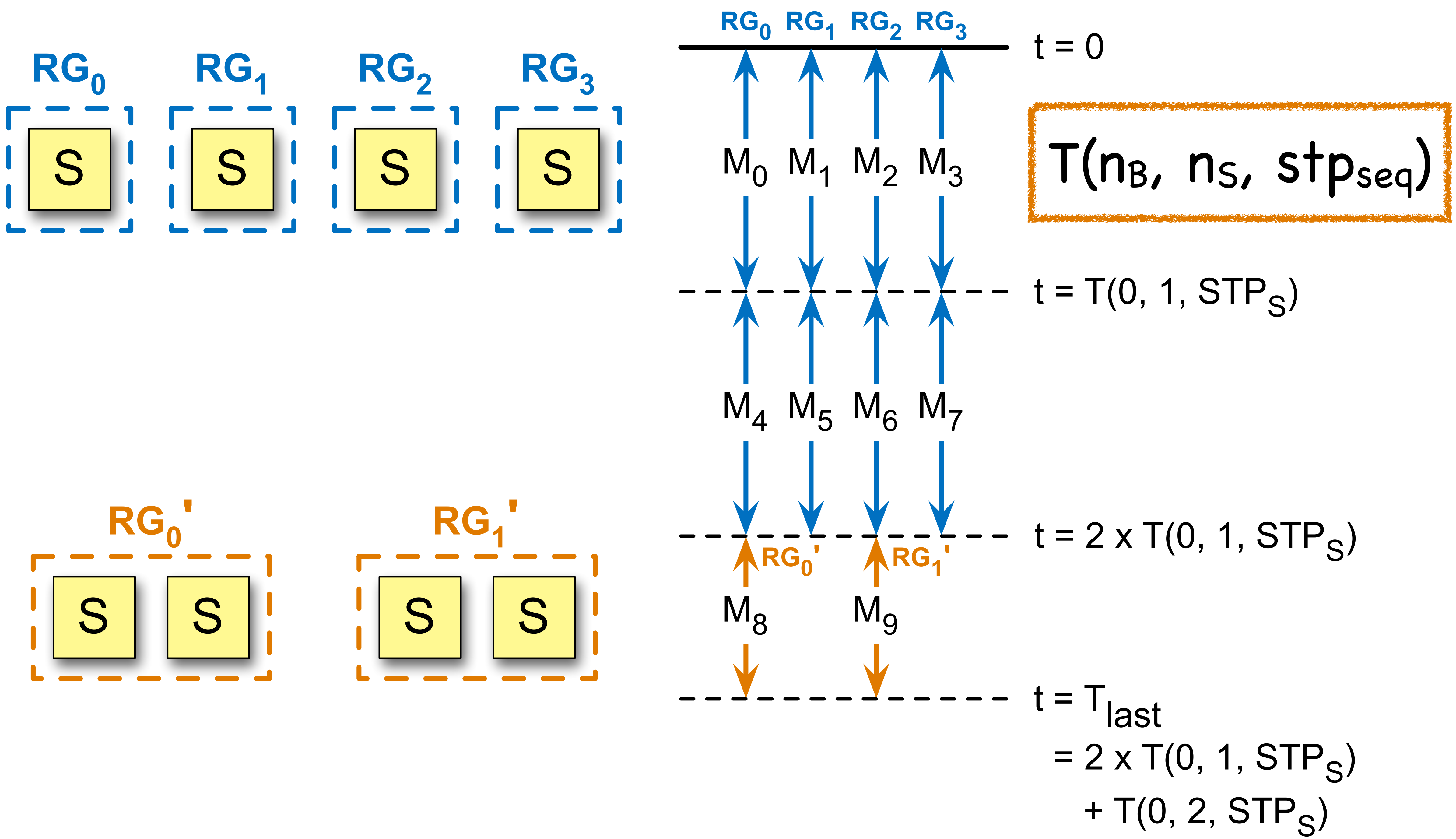
$$THP_{Asym} = \frac{1}{\boxed{\dfrac{(1-F)}{STP_B} + \dfrac{F}{STP_B + N_S \times STP_S}}}$$

$$T(n_B, n_S, stp_{seq}) = \frac{(1-F)}{stp_{seq}} + \frac{F}{(n_B \times STP_B) + (n_S \times STP_S)}$$

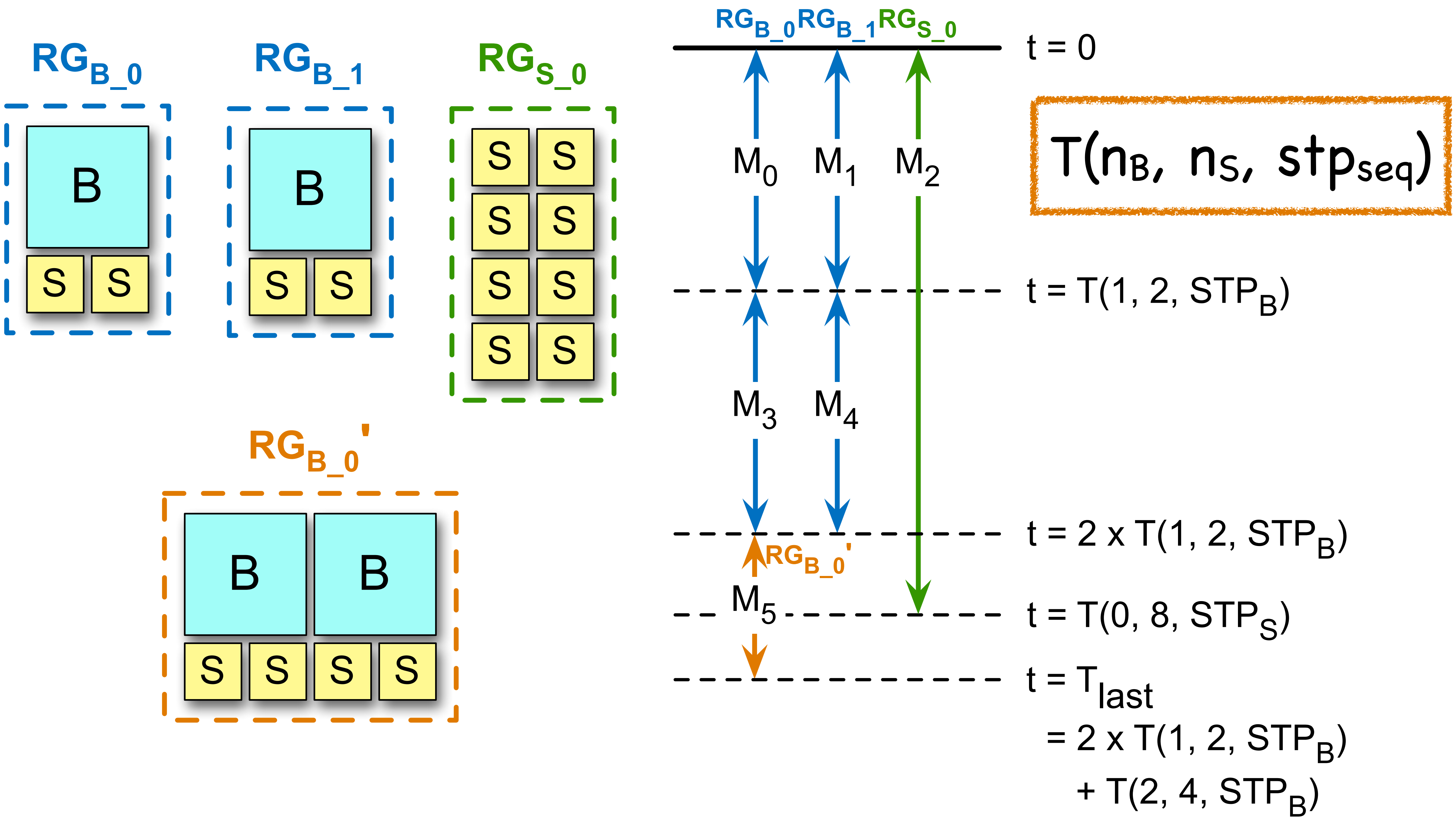
$$THP_{Symm} = \frac{1}{T(0, N_S, STP_S)}$$

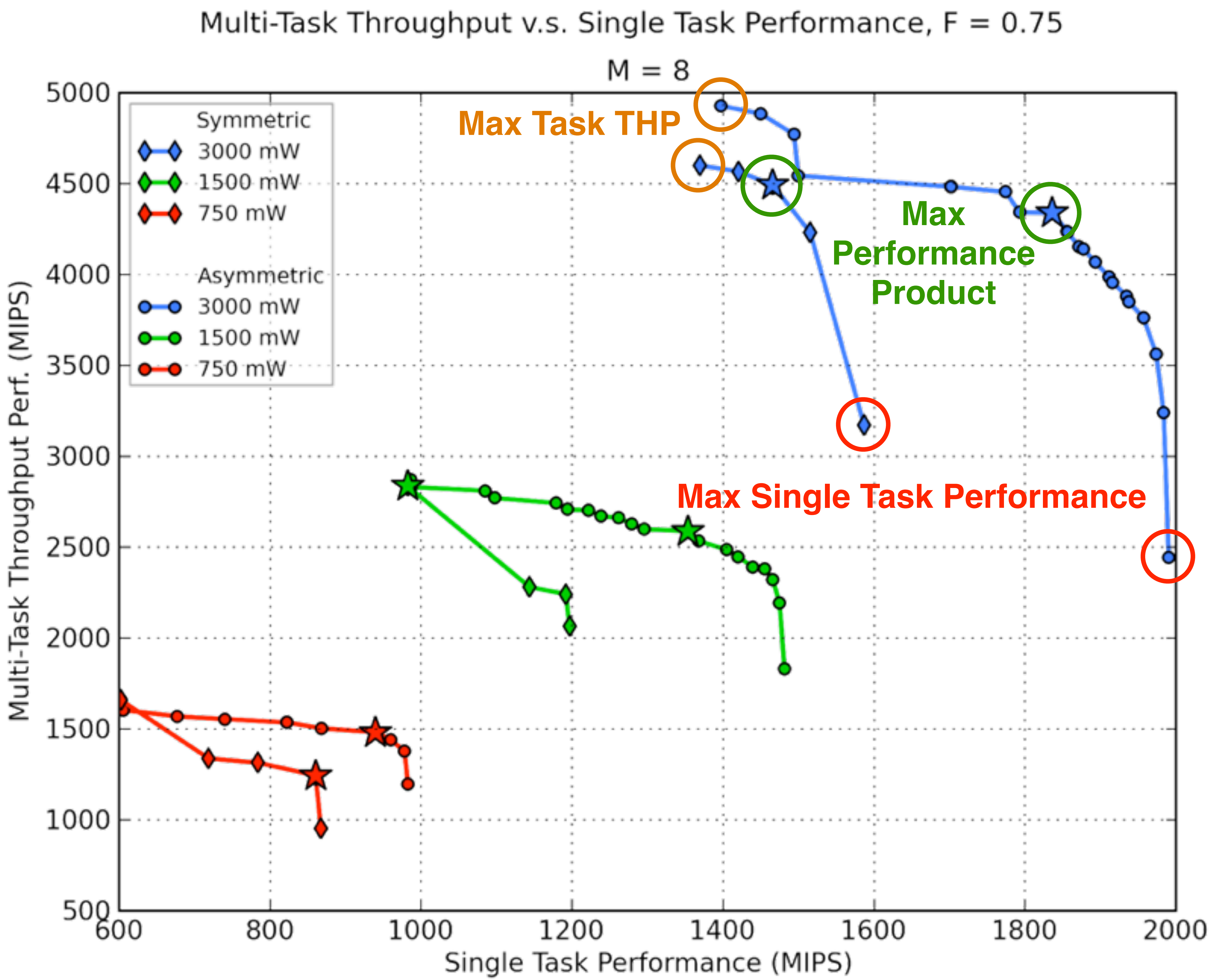
$$THP_{Asym} = \frac{1}{T(1, N_S, STP_B)}$$

Multi-Task Symmetric Model

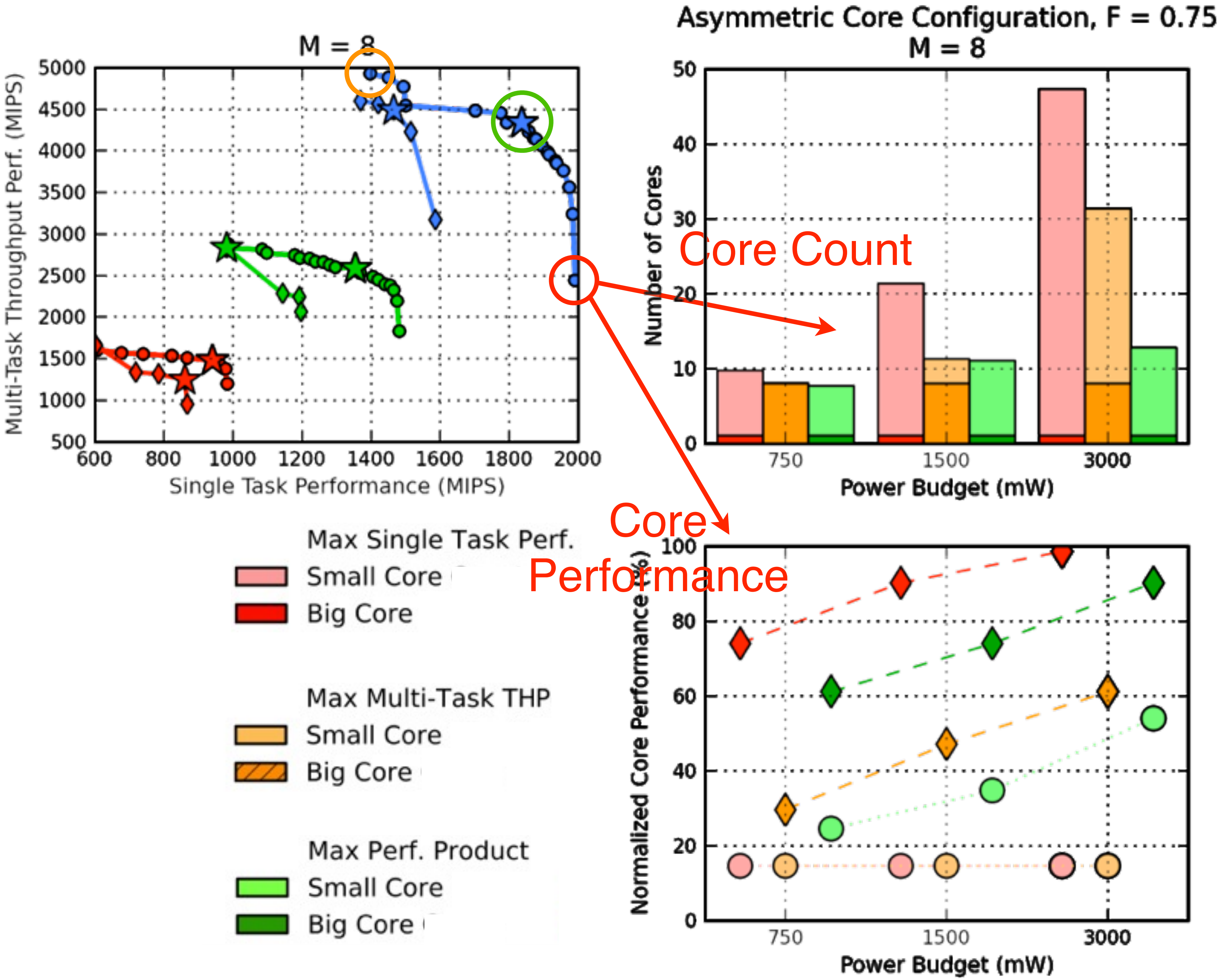


Multi-Task Asymmetric Model





Multi-Task Throughput v.s. Single Task Performance, F = 0.75
M = 8
Symmetric
3000 mW
1500 mW
750 mW
Asymmetric
3000 mW
1500 mW
750 mW
Max Task THP
Max Performance Product
Max Single Task Performance
Multi-Task Throughput Perf. (MIPS)
Single Task Performance (MIPS)
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
600
800
1000
1200
1400
1600
1800
2000


M = 8
Multi-Task Throughput Perf. (MIPS)
Single Task Performance (MIPS)
Asymmetric Core Configuration, F = 0.75
M = 8
Number of Cores
Power Budget (mW)
Core Count
Core
Performance
Normalized Core Performance (%)
Power Budget (mW)
Max Single Task Perf.
Small Core
Big Core
Max Multi-Task THP
Small Core
Big Core
Max Perf. Product
Small Core
Big Core

Takeaways

- Asymmetric design benefits more when power budget is higher
- Core configurations varied significantly for different performance metrics
- Almost every size of single cores is selected to form an optimized configuration

Conclusion

- For future processors, improving both single thread performance and throughput performance is significant
- We construct an analytical model that computes both type of performance under a power constraint for different multicore architecture
- Future work: considering overheads like cache sharing and NoC, extending the model for heterogeneous multi-task, and analyzing QoS metrics

Questions?