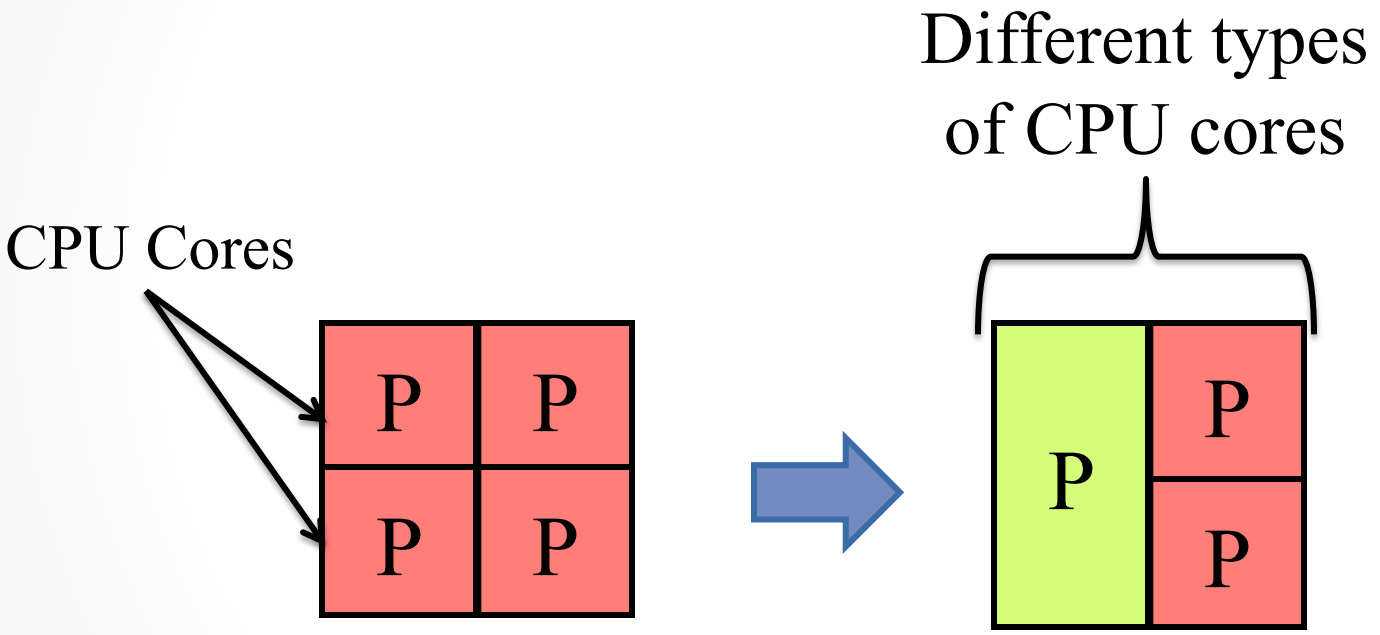
Analyzing Performance Asymmetric Multicore Processors for Latency Sensitive Datacenter Applications

HotPower'10
October 3rd, 2010

Vishal Gupta* (Georgia Tech)
Ripal Nathuji (Microsoft Research)

* Work done during summer internship at Microsoft Research

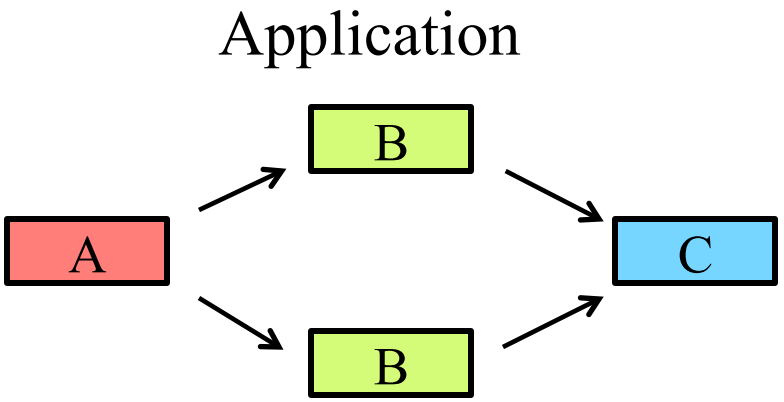
What are AMPs?

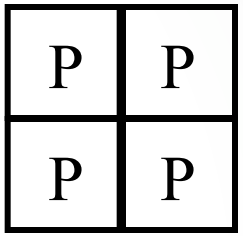


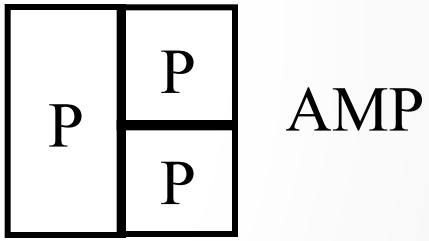
Symmetric
multicore processor
SMP

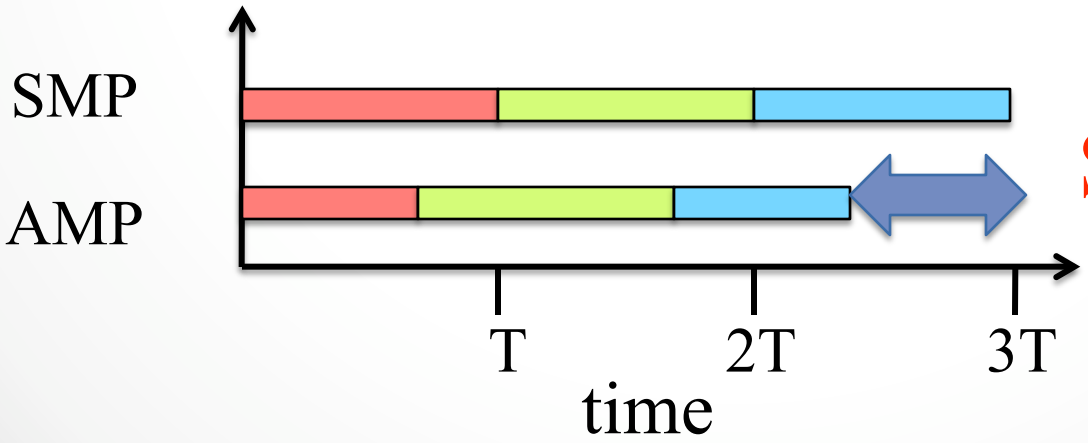
Asymmetric
multicore processor
AMP

Why AMPs?







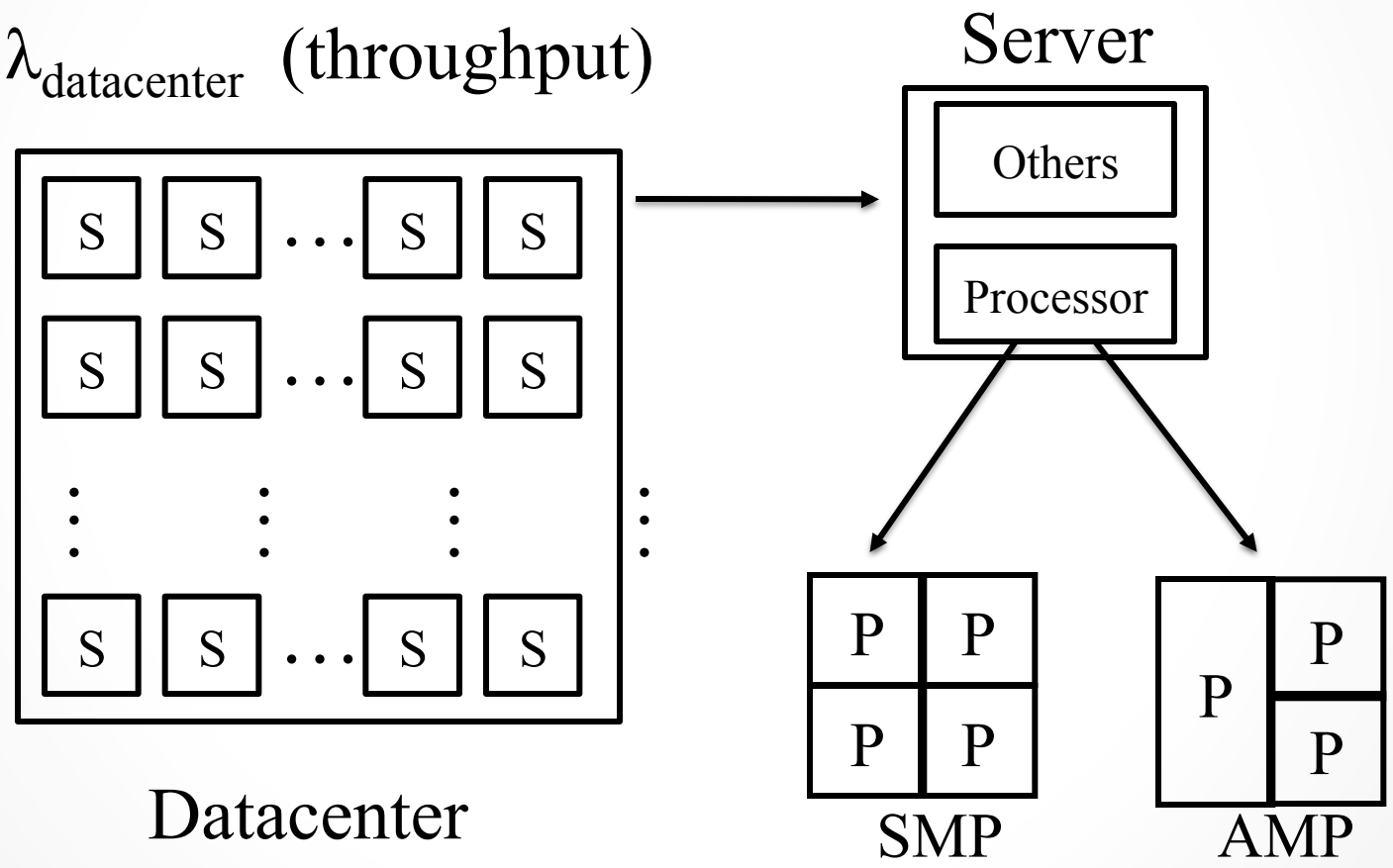


Speedup!

Goals

- How good are AMPs as compared to SMPs?
- Can datacenter applications save power using AMPs?

Datacenter Model



Objective

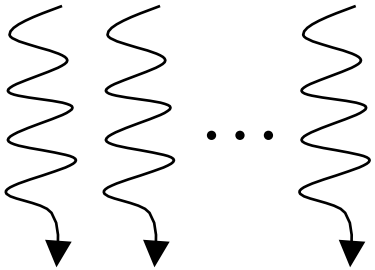
$$P^{AMP}_{datacenter} < P^{SMP}_{datacenter} \; ?$$

- Constant work
- Meet latency SLA

AMP Use Cases

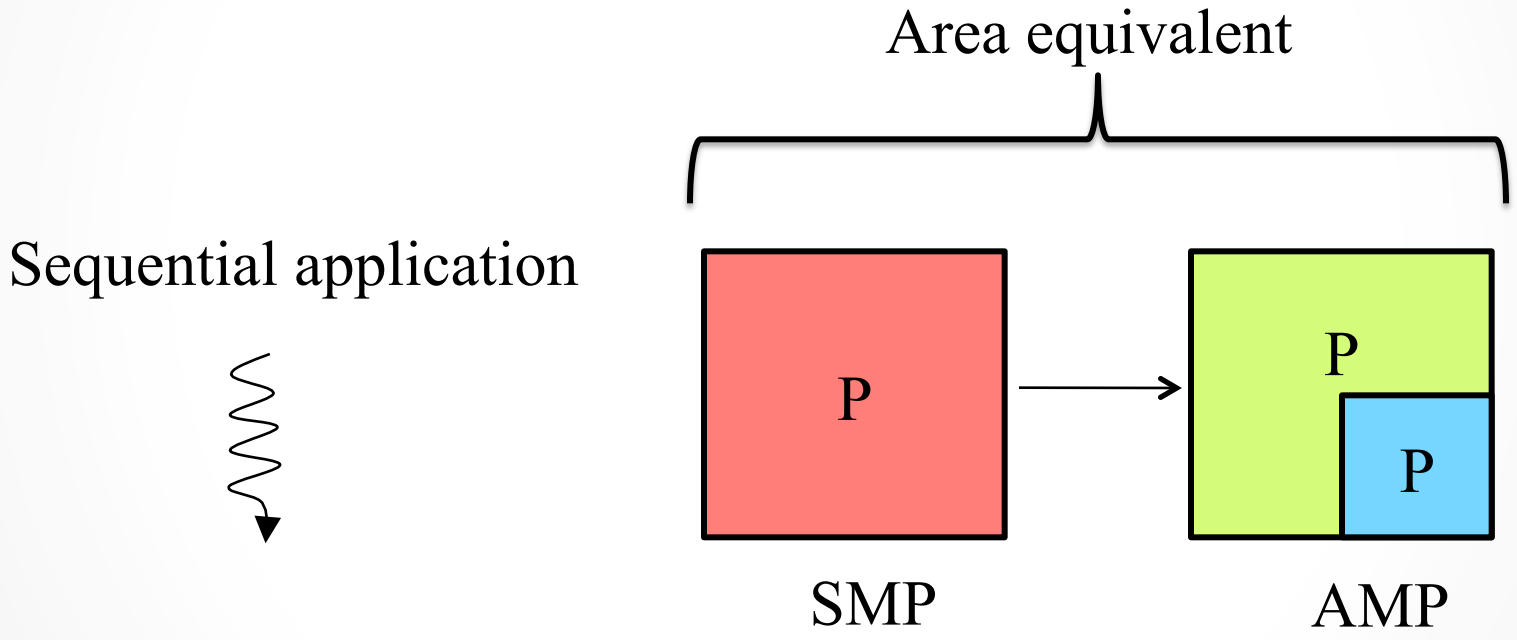
- Energy Scaling
- Parallel Speedup

Sequential execution

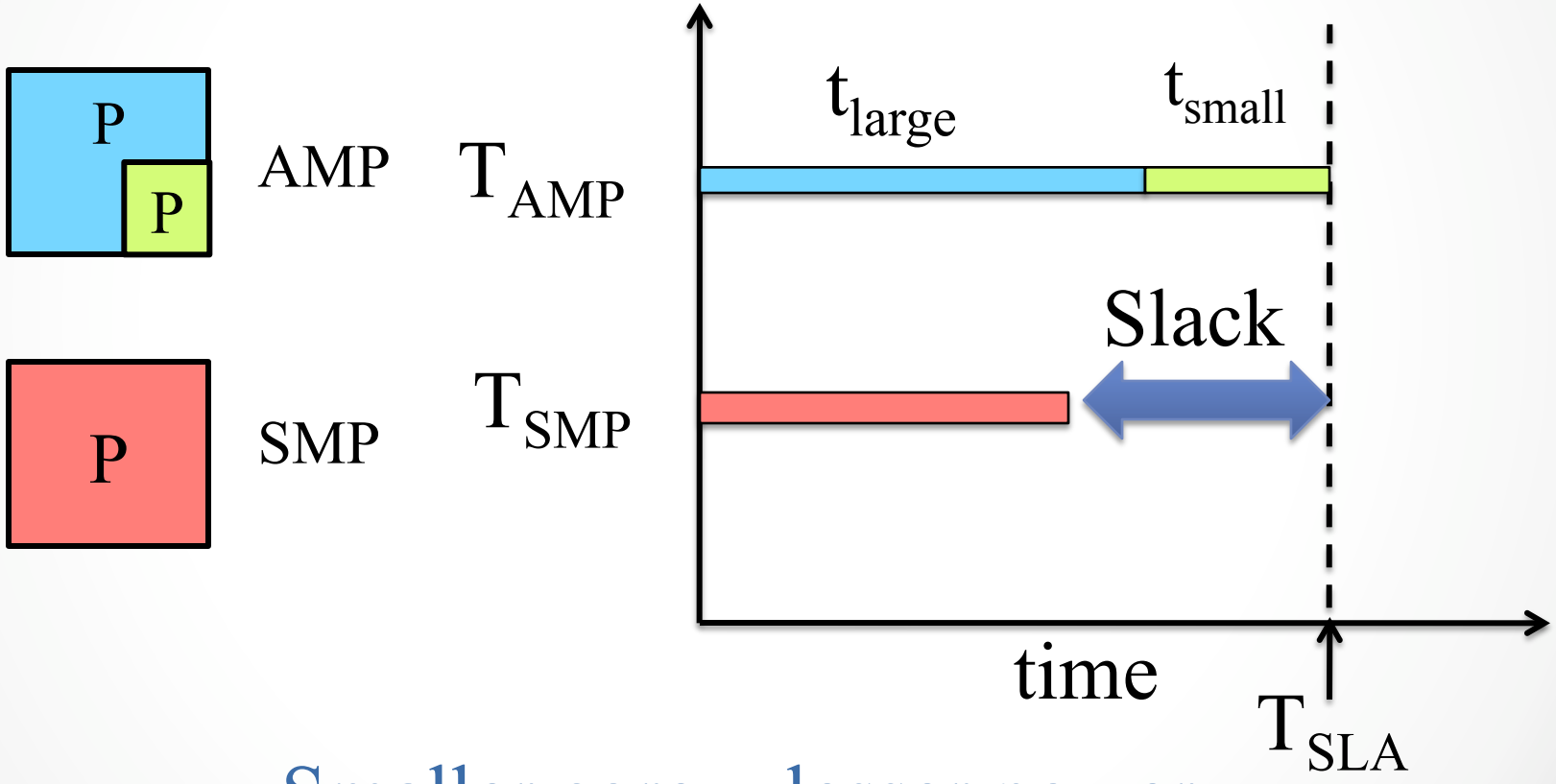


Parallel execution

Energy Scaling (ES)

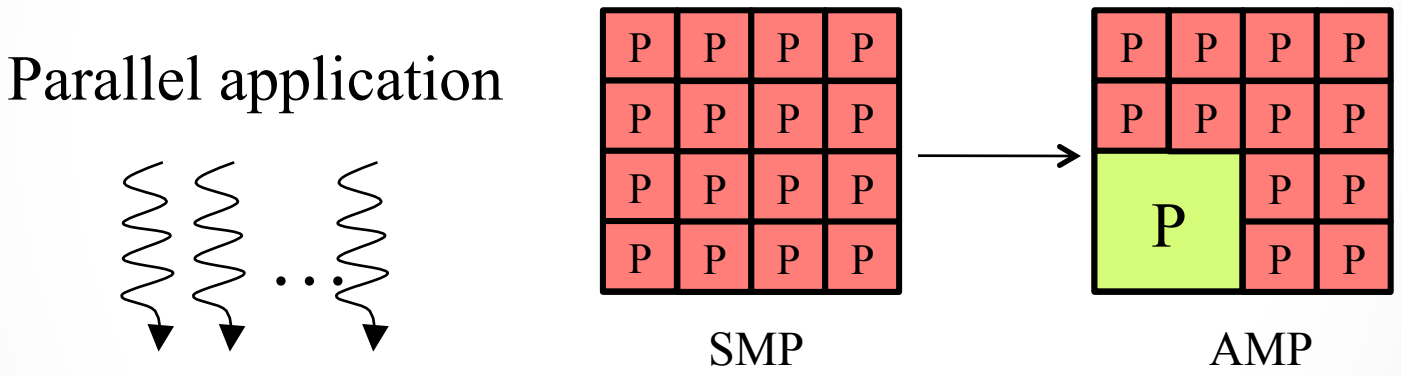


Energy Scaling (ES)

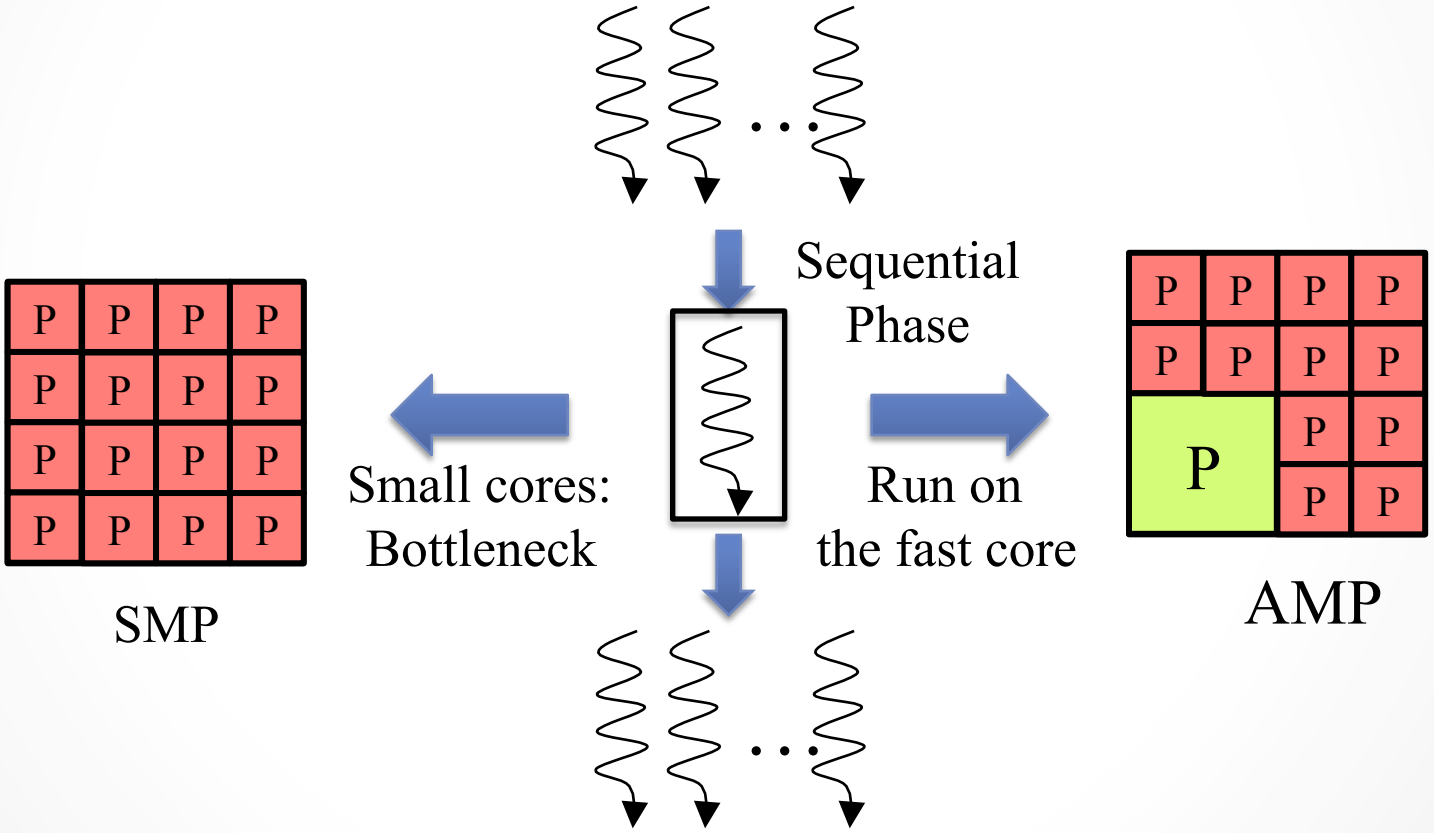


Smaller core = lesser power

Parallel Speedup (PS)

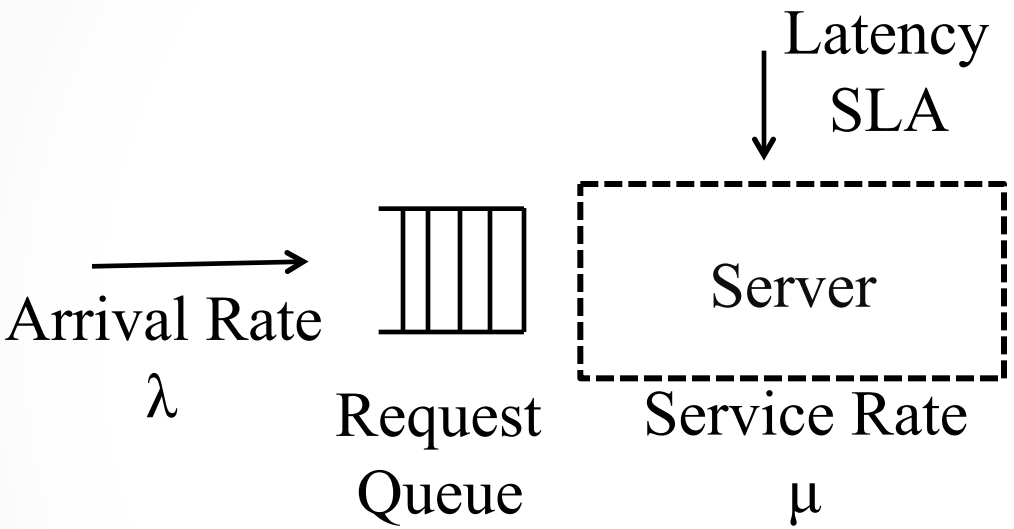


Parallel Speedup (PS)



Speedup = Higher throughput

Queuing Model for Server



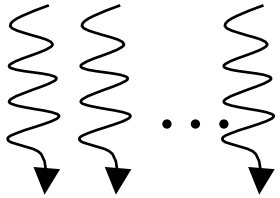
M/M/1 Queuing Model

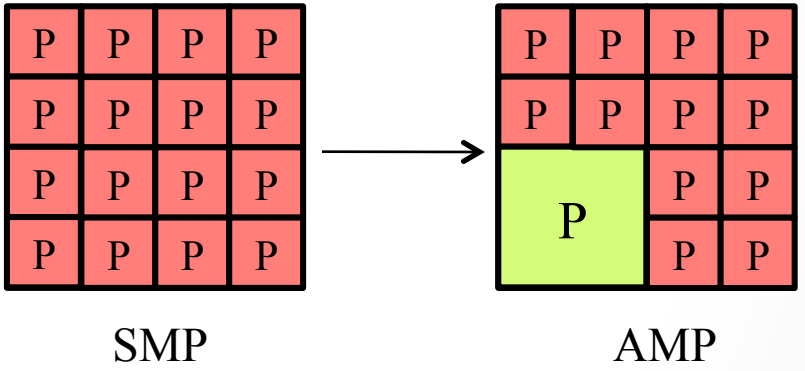
Avg. Response Time $E[T] = \frac{1}{\mu - \lambda}$

Modeling Service Rate (μ)

Parallel Speedup (PS)
(refer to paper for ES)

Parallel application



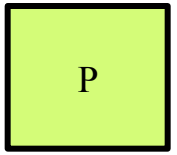


Amdahl's Law for Multicores

Amdahl's Law for Multicores

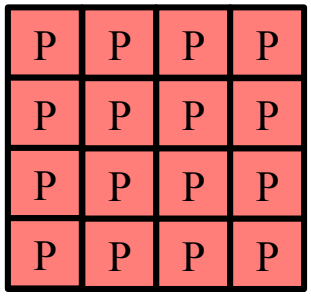
r = Area(Big/Core)

Area = 1

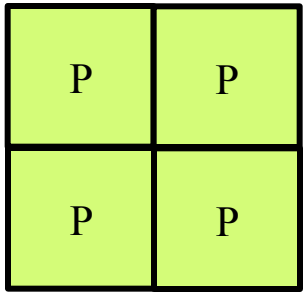


Area = r
Perf = perf(r)

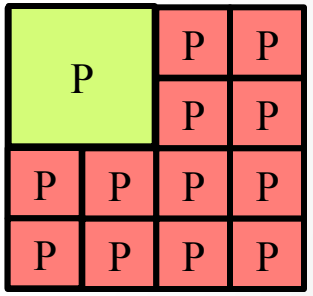
n = Chip area



SMP
n=16, r=1



SMP
n=16, r=4



AMP
n=16, r=4

f = fraction of computation that can be parallelized

Amdahl's Law for Multicores

$$\mu_{SMP}(f,n,r) = \frac{1}{\frac{1-f}{perf(r)} + \frac{f}{\frac{n}{r} * perf(r)}}$$

$$\mu_{AMP}(f,n,r) = \frac{1}{\frac{1-f}{perf(r)} + \frac{f}{n-r}}$$

Ref: Hill and Marty, Amdahl's law in the multicore era (IEEE Computer'08)

Server Throughput (λ)

$$\lambda_{server}^{peak} = \mu - \frac{1}{T_{SLA}}$$

Datacenter capacity =
No. of servers * Server throughput

Constant Work

$$\lambda_{datacenter} = N_{server}^{SMP} * \lambda_{server}^{SMP}$$

$$\lambda_{datacenter} = N_{server}^{AMP} * \lambda_{server}^{AMP}$$

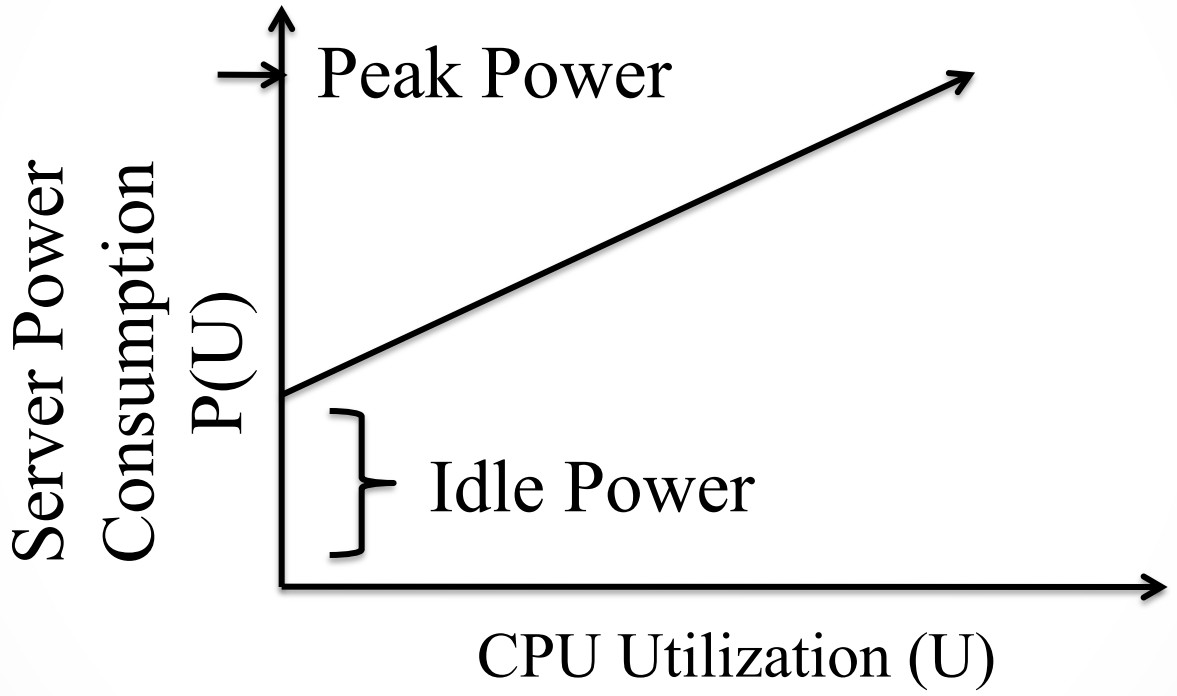
Datacenter Power Consumption

Datacenter power (P) =
No. of servers * Server power

$$P_{datacenter}^{SMP} = N_{server}^{SMP} * P_{server}^{SMP}$$

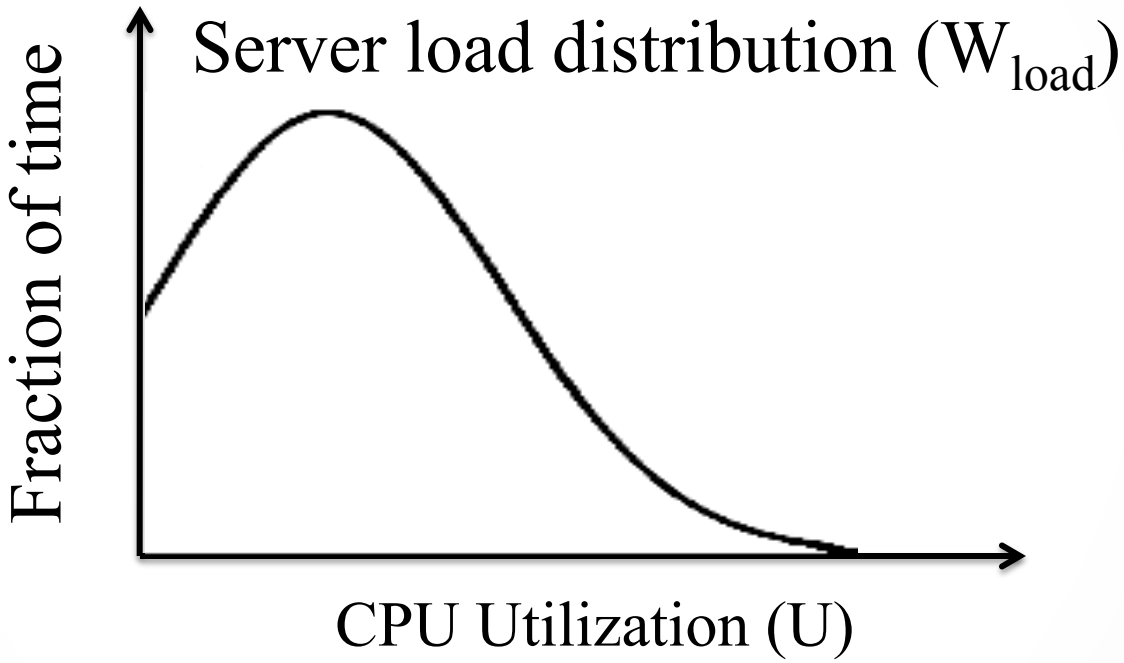
$$P_{datacenter}^{AMP} = N_{server}^{AMP} * P_{server}^{AMP}$$

Modeling Server Power



Ref: The Case for Energy-Proportional Computing, Barroso & Hölzle, IEEE Computer 2007

Modeling Server Power



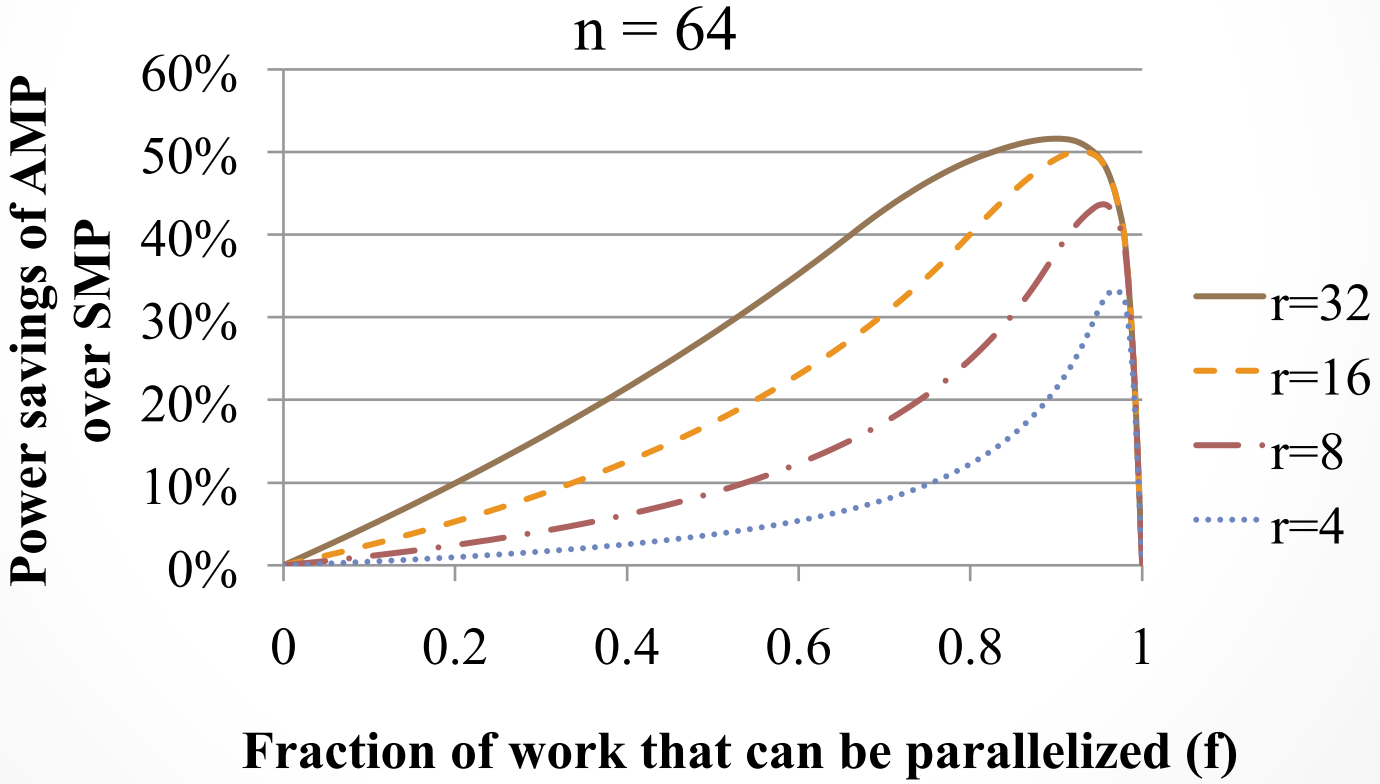
$$P_{server} = \sum W_{load}(U) * P_{server}(U)$$

Results

$$P_{datacenter}^{AMP} < P_{datacenter}^{SMP} ?$$

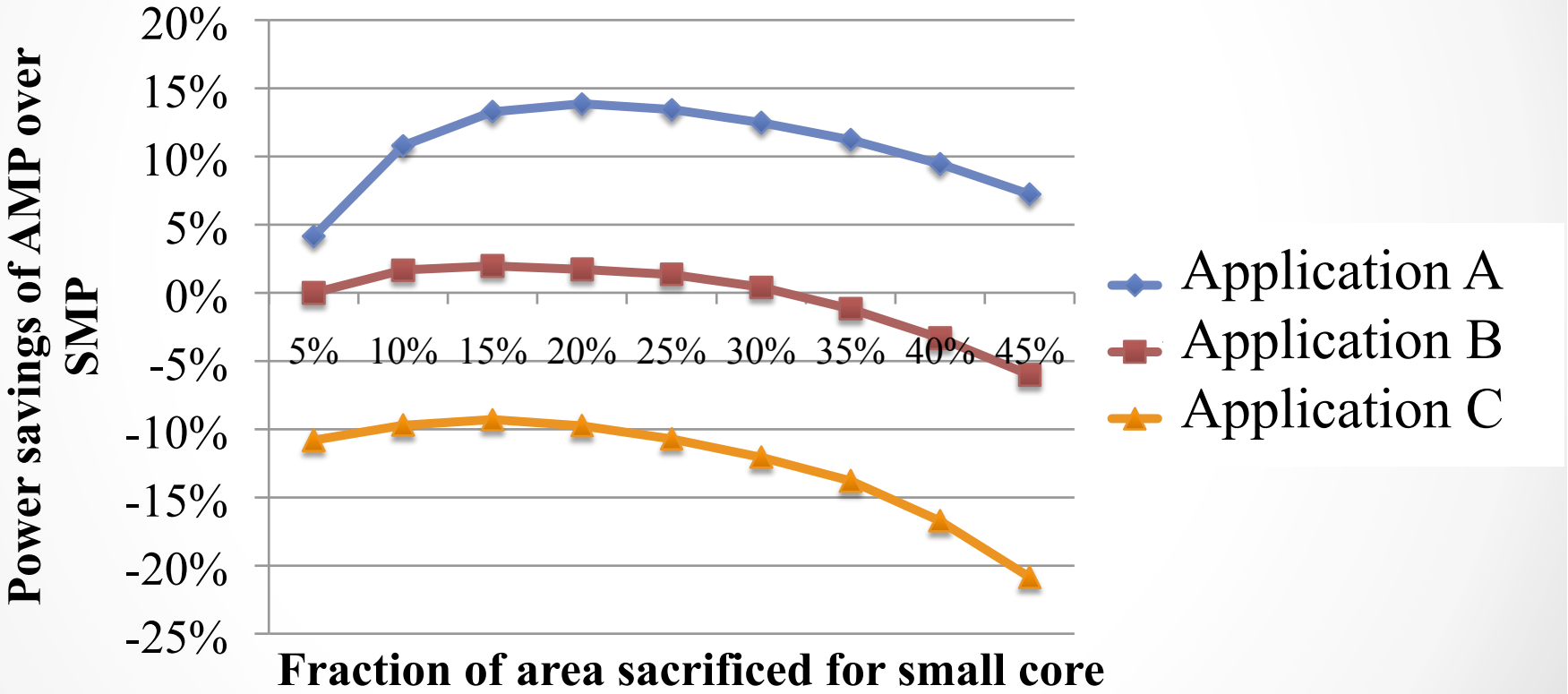
PS: Power Savings

Upto 52% power savings



ES: Power Saving

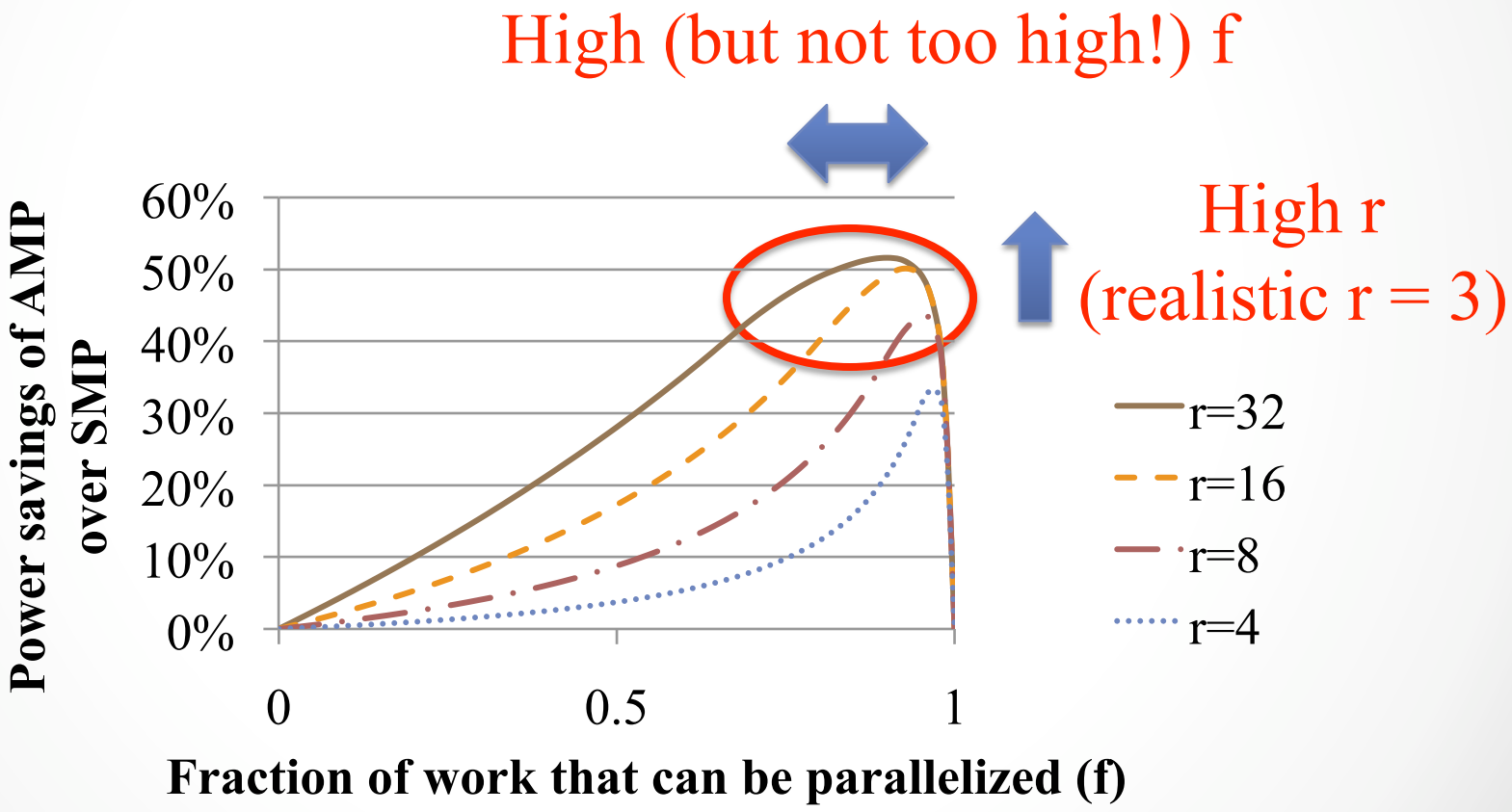
Upto 14% power savings



AMP Power Savings

- PS looks more promising that ES
- Can we achieve these savings in reality?

H/W & S/W Design Challenges



Practical Considerations

- **Scalability**: Amdahl's law assumes unbounded scalability
- **Migration overhead**: zero migration overhead
- **Perfect scheduling**: oracle scheduler

Actual savings are going to be lower

Conclusions

- Potential for power savings in datacenters using AMPs
- Parallel Speedup more promising than Energy Scaling
- Practical considerations to realize full benefits

Future work:

Extend our analysis to functional asymmetry