

# **PARM** POWER-AWARE RESOURCE MANAGER

Maximizing Data Center Performance Under Strict Power Budget

*Osman Sarood, **Akhil Langer\***, Abhishek Gupta, Laxmikant Kale*

*Parallel Programming Laboratory  
Department of Computer Science  
University of Illinois at Urbana-Champaign*

*19<sup>th</sup> November*

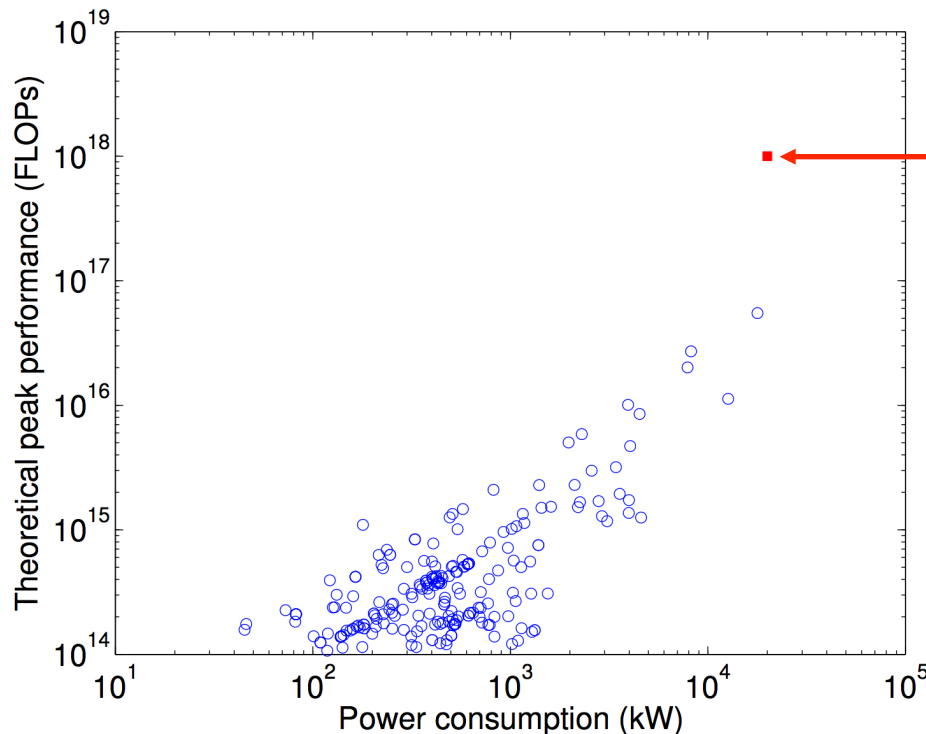
*BoF: Dynamic Power Management for MW-sized Supercomputer Centers  
Supercomputing 2014*

*New Orleans, LA*



# Major Challenge to Achieve Exascale

Power consumption for Top500



Exascale in 20MW!

# Data Center Power

How is power demand of data center calculated?

- ❑ Using Thermal Design Power (TDP)!

However, TDP is hardly reached!!

## Constraining CPU/Memory power



*Intel Sandy Bridge*

- ❑ Running Average Power Limit (RAPL) library
  - measure and set CPU/memory power

# Constraining CPU/Memory power



## *Intel Sandy Bridge*

- ❑ Running Average Power Limit (RAPL) library
  - measure and set CPU/memory power

## *Solution to Data Center Power*

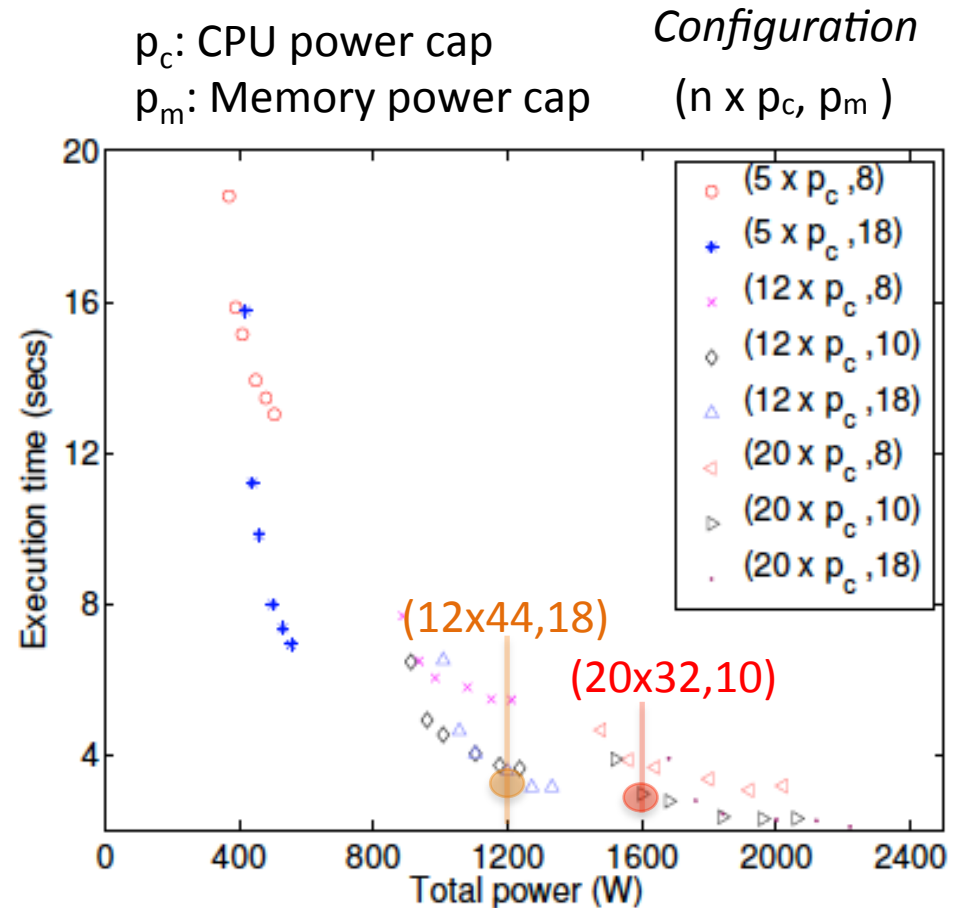
- ❑ Constrain power consumption of nodes
- ❑ *Overprovisioning* - Use more nodes than conventional data center for same power budget

# Application Performance with Power

Application performance does not improve proportionately with increase in power cap

Run on larger number of nodes each capped at lower power level

Performance of LULESH at different configurations



# Problem Statement

Maximizing Data Center Performance Under Strict Power Budget

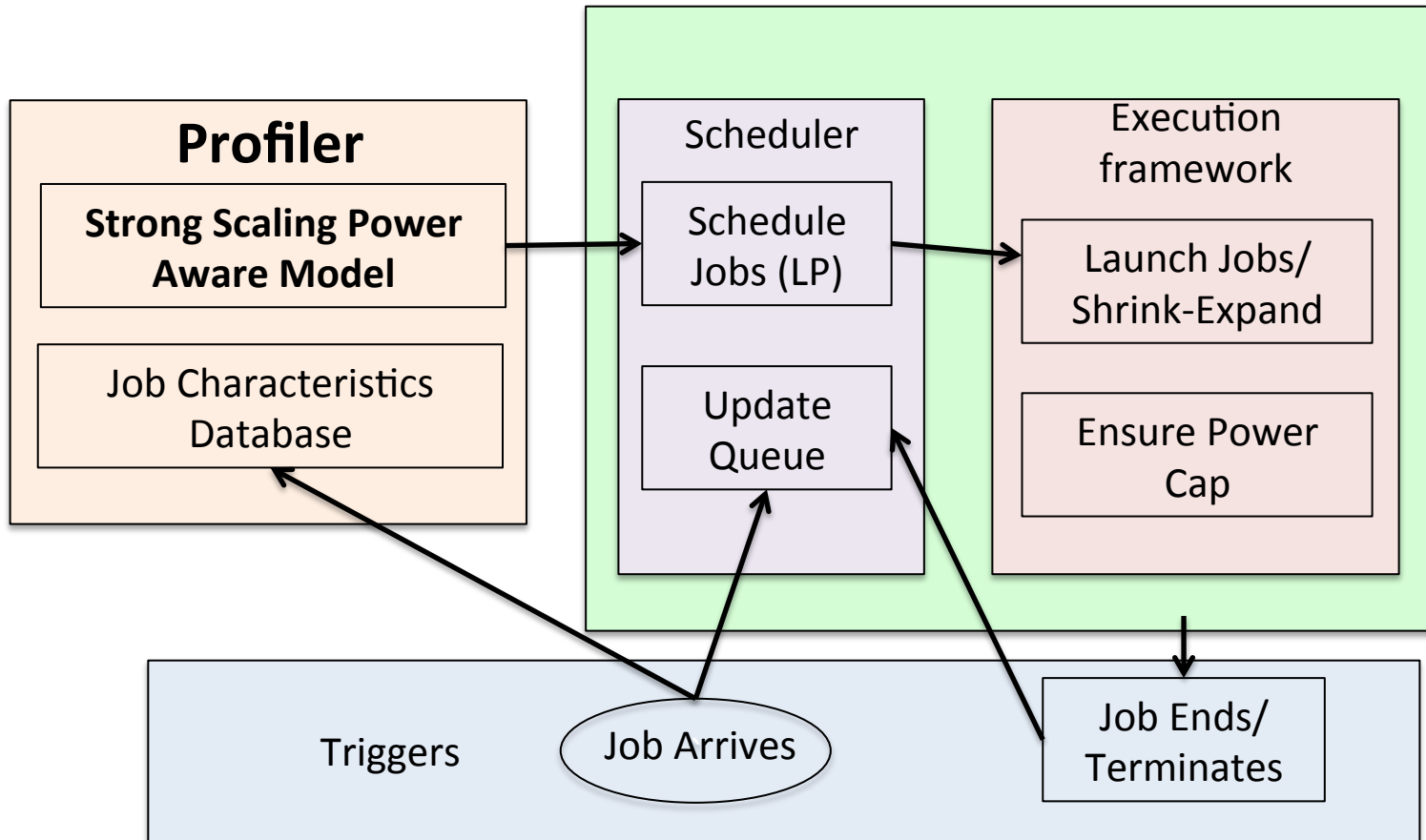
*Data center capabilities*

- Power capping ability
- Overprovisioning

*Job features (Optional)*

- Moldability
- Malleability
  - Charm++

# PARM POWER-AWARE RESOURCE MANAGER



# PASS Power Aware Strong Scaling Model

Execution time as a function of power and number of nodes

## Time vs Scale

Downey's strong scaling

$$t = F(n, A, \sigma)$$

- $n$ : number of nodes
- $A$ : Average Parallelism
- $\sigma$ : duration of parallelism  $A$

## Time vs Frequency

$$t(f) = \begin{cases} \frac{W_{cpu}}{f} + T_{mem}, & \text{for } f < f_h \\ T_h, & \text{for } f \geq f_h \end{cases}$$

- $W_{cpu}$ : CPU work
- $T_{mem}$ : memory work
- $T_h$ : minimum exec time

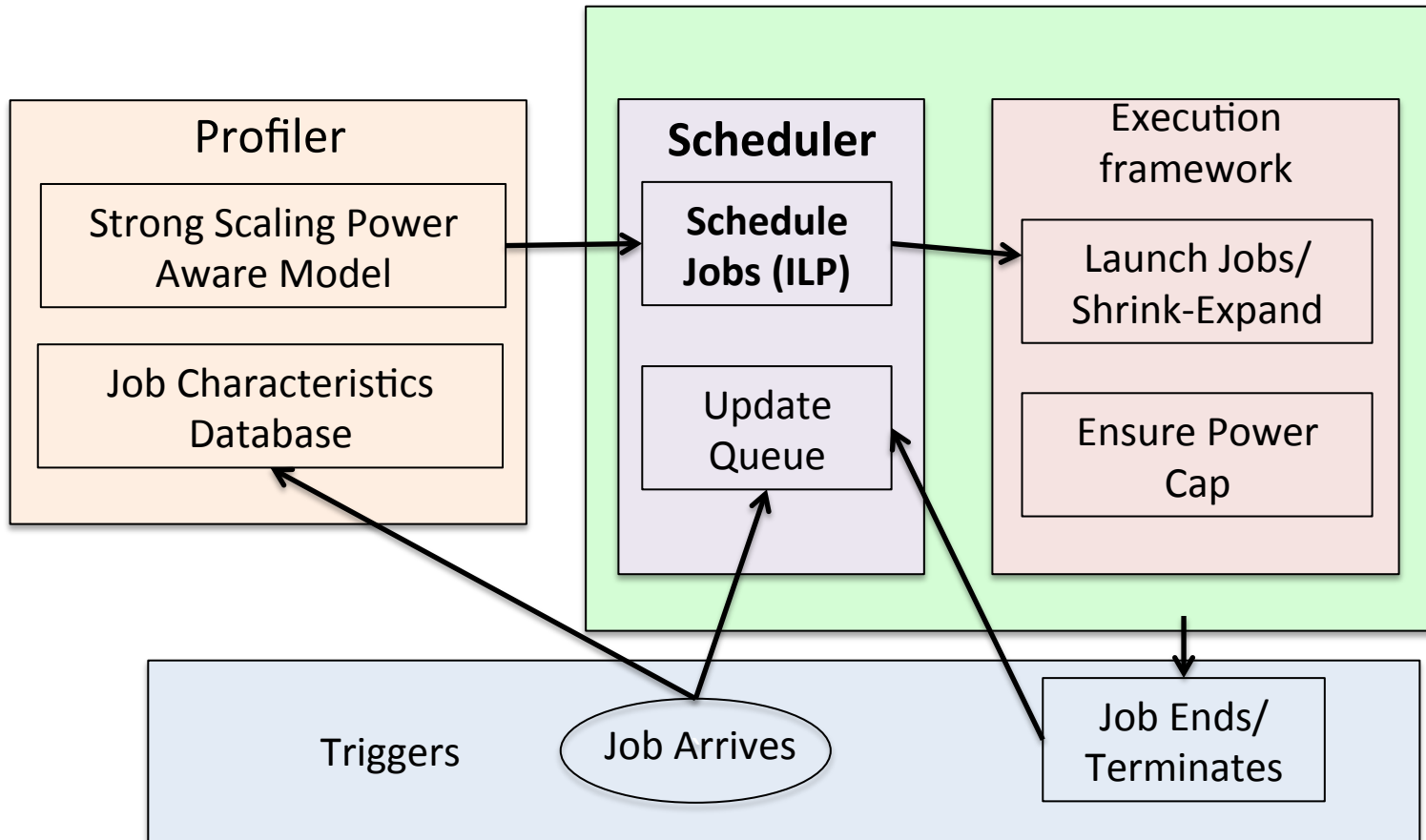
## Power vs Frequency

$$p = p_{core} + \sum_{i=1}^3 g_i L_i + g_m M + p_{base}$$

- $p_{core}$ : core power
- $g_i$ : cost level  $i$  cache access
- $L_i$ : #level  $i$  accesses
- $g_m$ : cost of mem access
- $M$ : #mem accesses
- $p_{base}$ : idle power



# PARM POWER-AWARE RESOURCE MANAGER



# Scheduler - Integer Linear Program (ILP) Formulation

*Objective Function*

$$\sum_{j \in \mathcal{J}} \sum_{n \in N_j} \sum_{p \in P_j} w_j * s_{j,n,p} * x_{j,n,p}$$

*Select One Resource Combination Per Job*

$$\sum_{n \in N_j} \sum_{p \in P_j} x_{j,n,p} \leq 1 \quad \forall j \in \mathcal{I}$$

$$\sum_{n \in N_j} \sum_{p \in P_j} x_{j,n,p} = 1 \quad \forall j \in \mathcal{I}$$

*Bounding total nodes*

$$\sum_{j \in \mathcal{J}} \sum_{p \in P_j} \sum_{n \in N_j} n x_{j,n,p} \leq \mathbf{N}$$

*Bounding power consumption*

$$\sum_{j \in \mathcal{J}} \sum_{n \in N_j} \sum_{p \in P_j} (n * (p + W_{base})) x_{j,n,p} \leq W_{max}$$

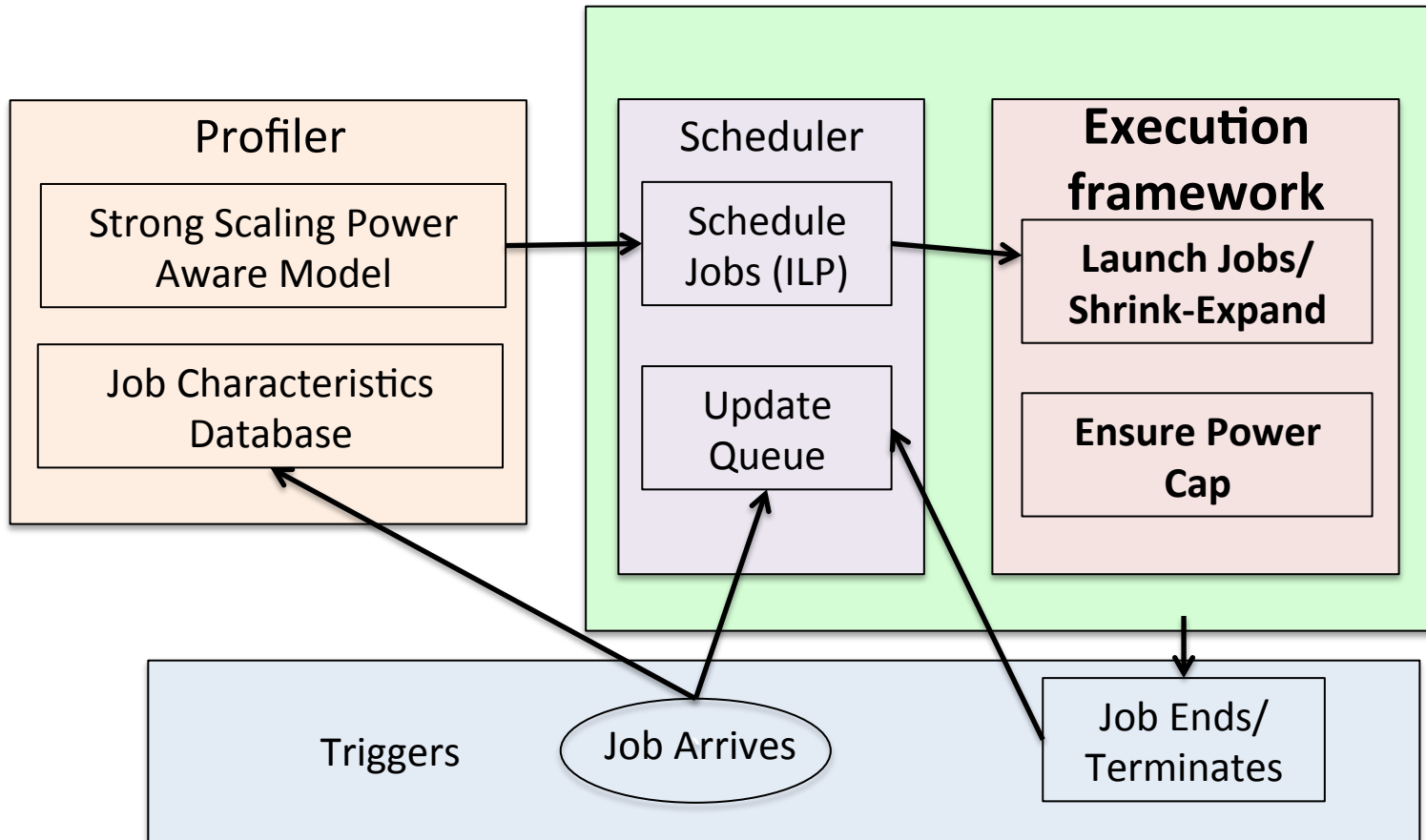
*Disable Malleability (Optional)*

$$\sum_{n \in N_j} \sum_{p \in P_j} n x_{j,n,p} = n_j \quad \forall j \in \mathcal{I}$$

Maximizing throughput  
makes online ILP  
optimization intractable,  
instead

maximize sum of  
power-aware speedup of  
selected jobs

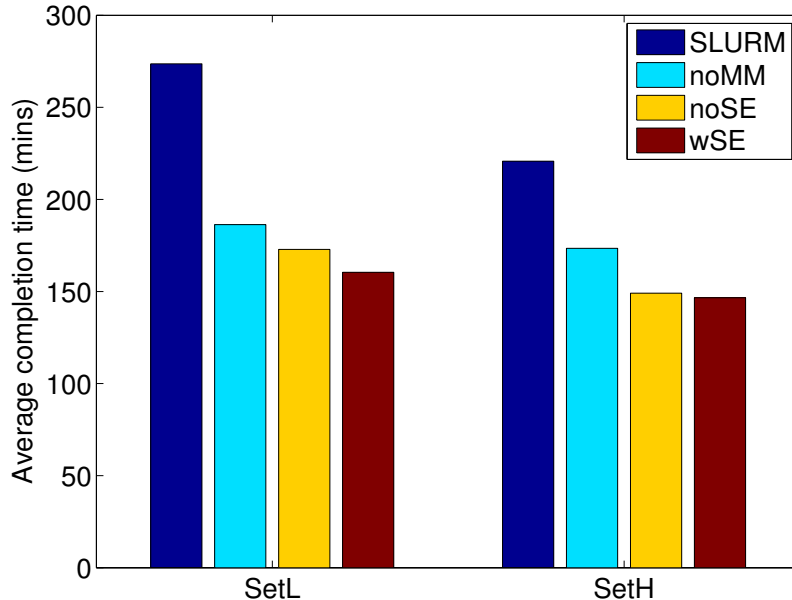
# PARM POWER-AWARE RESOURCE MANAGER



# Performance Results

Lulesh, AMR, LeanMD, Jacobi and Wave2D

38-node Intel Sandy Bridge Cluster, 3000W budget



## Description

- noMM**: without Malleability and Moldability
- noSE**: with Moldability but no Malleability
- wSE**: with Moldability and Malleability

1.7X improvement in throughput

# Takeaways

Significant improvement in throughputs

- Power-aware characteristics (PASS model)
- CPU power capping
- Overprovisioning

Sophisticated ILP scheduling methodology useful for resource assignment

Adaptive runtime system further increases benefits by allowing malleability

*Talk: Maximizing Throughput of Overprovisioned HPC Data centers Under a Strict Power Budget. SC '14.*

*Thursday (tomorrow) 10:30 am, 393-94-95.*

# THANK YOU!

## **PARM** POWER-AWARE RESOURCE MANAGER

Maximizing Data Center Performance Under Strict Power Budget

*Osman Sarood, **Akhil Langer\***, Abhishek Gupta, Laxmikant Kale*

*Parallel Programming Laboratory  
Department of Computer Science  
University of Illinois at Urbana-Champaign*

*19<sup>th</sup> November  
Supercomputing 2014  
New Orleans, LA*

