Parallel Application Power and Performance Prediction Modeling Using Simulation

Kishwar Ahmed, Kazutomo Yoshii*, Samia Tasnim**

University of South Carolina Beaufort

*Argonne National Laboratory

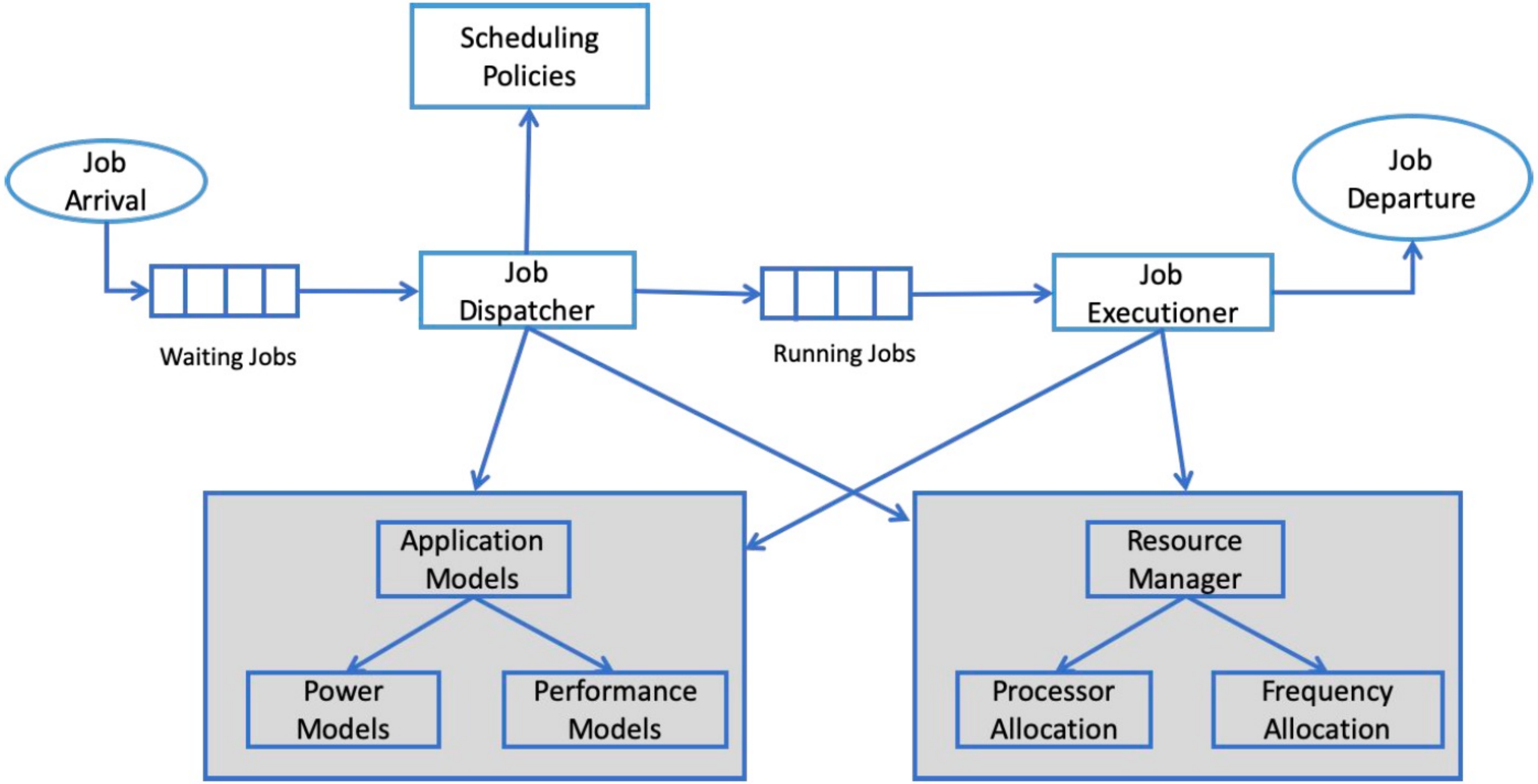
**Florida A&M University

Winter Simulation Conference (WSC 2021), Phoenix, AZ, USA

Contributions

- Power and performance prediction model of HPC applications
  - Different power-capping values
- Heterogeneous computing architecture
- Based on parallel discrete event simulation
  - Simulus (https://simulus.readthedocs.io/en/latest/)
- Simulation case studies

Overall model



Power prediction model

- Power consumption at power-cap, P_j :

$$p_j = a + b \cdot P_j + c \cdot {P_j}^2 + d \cdot {P_j}^3$$

- Power-capping constraint:

$$P_{min} \leq P_j \leq P_{max}$$

Resource allocation model

- Determine optimal power-cap values:

$$\text{Minimize } \sum_{j=1}^{n_j} e_j$$

$$\text{Subject to } \sum_{j \epsilon R} n_j \leq \hat{n}$$

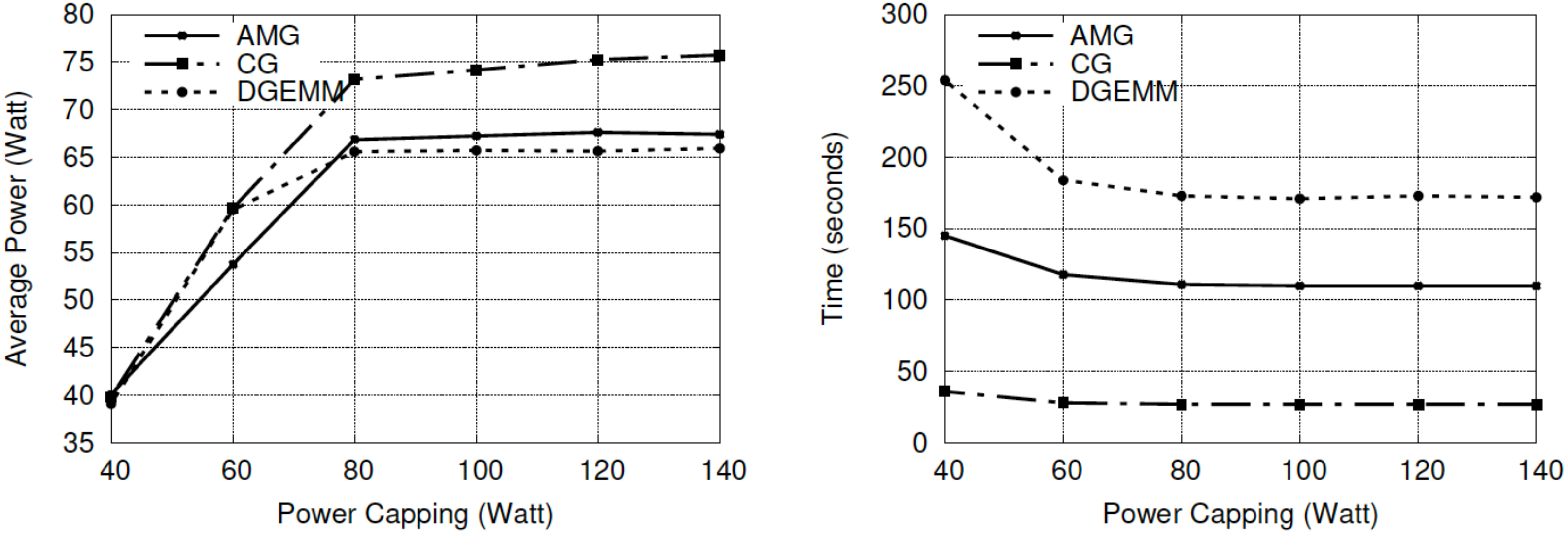
$$p_j + \sum_{i \epsilon R} p_i \leq \hat{p}$$

Case studies

- Case study#1: power and performance impact on CPU node
- Case study#2: power and performance impact on GPU node
- Case study#3: power and performance impact on FPGA node
- Case study#4: optimal power allocation

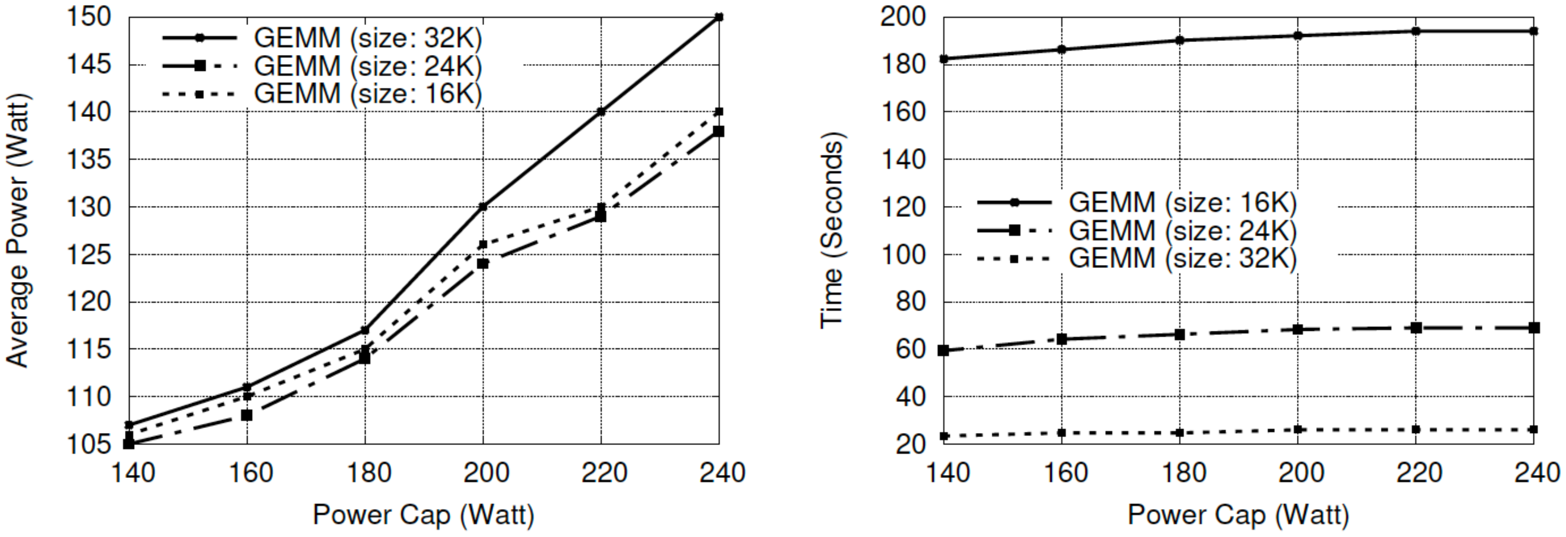
Case study#1

- Power measurement on CPU node
- Used pycoolr to change power-cap limit



Case study#2

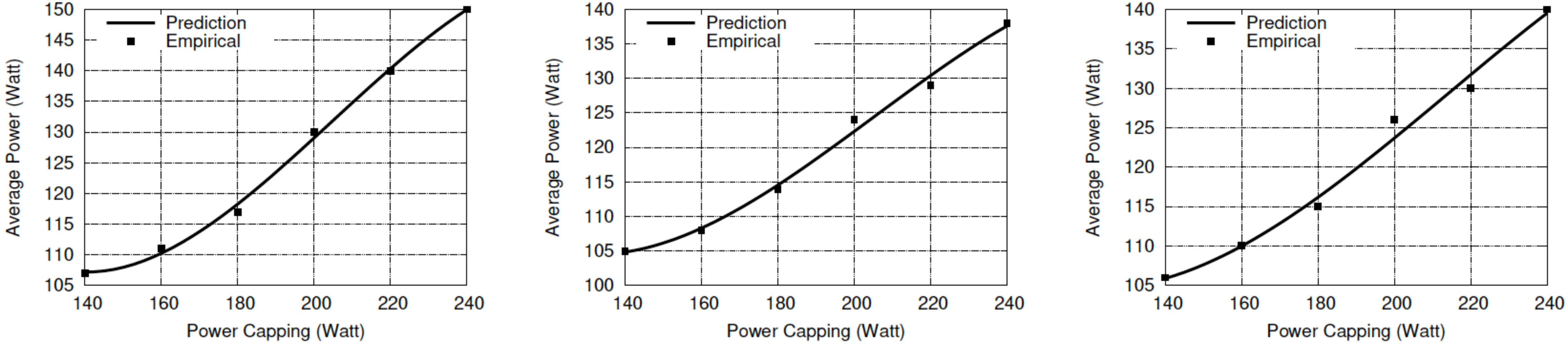
- Power measurement on GPU node
- Existing study (Krzywaniak and Czarnul 2019)



Krzywaniak, A., and P. Czarnul. 2019. “Performance/energy aware optimization of parallel applications on gpus under power capping”. In Proceedings of the International Conference on Parallel Processing and Applied Mathematics, 123–133. Springer.

Case study#2 (contd.)

- Prediction model to predict application characteristics



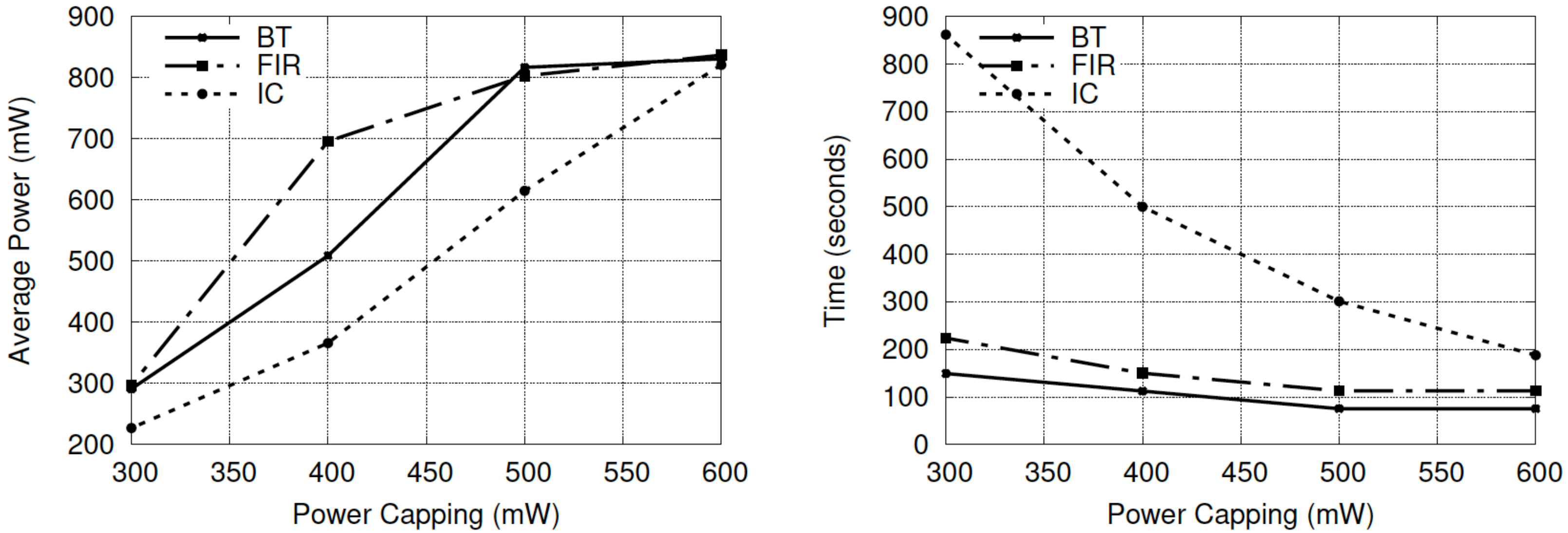
(a) GEMM (32K)

(b) GEMM (24K)

(c) GEMM (16K)

Case study#3

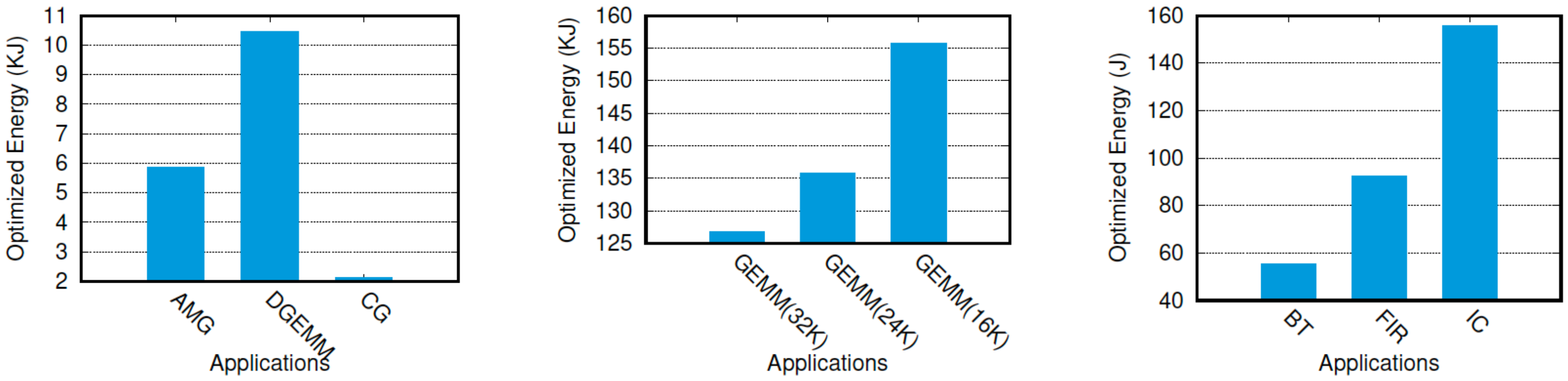
- Power measurement on FPGA node
- Existing study (Wu et al. 2016)



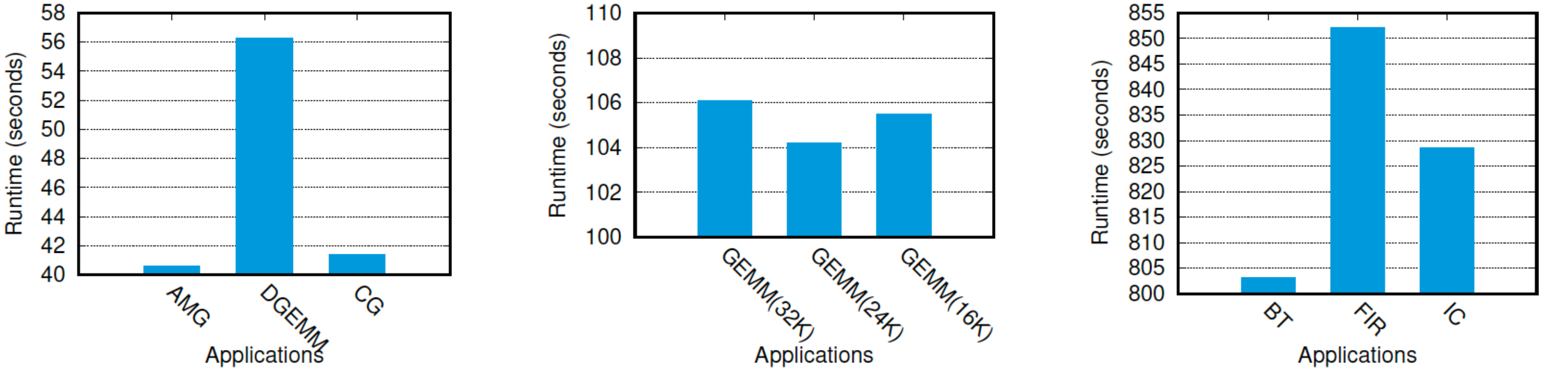
Wu, Y., D. S. Nikolopoulos, and R. Woods. 2016. “Runtime Support for Adaptive Power Capping on Heterogeneous SOCs”. In Proceedings of the 2016 International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation (SAMOS)

Case study#4

- Optimal resource allocation



(a) CPU (b) GPU (c) FPGA



(a) CPU (b) GPU (c) FPGA

Conclusions

- Power and performance prediction modeling of HPC applications
- Simulation using parallel discrete-event simulation engine
- Separate cases studies for CPUs, GPUs, and FPGAs
- Prediction model without a-priori application knowledge

Thank you!

Questions?: ahmedk@uscb.edu

Acknowledgements:

NSF

UNIVERSITY OF
South Carolina
Office of Research