The Curious Case of GPU Workload Analysis

- **Era of H/W Accelerators**
  - End of Moore’s law -> beginning of Accelerator innovations like GPUs, TPUs, and FPGAs.
  - State-of-the-art datacenter schedulers are still agnostic or at early stages to support GPUs.
  - Need for GPU-aware scheduling policies in datacenter orchestrators like Kubernetes.

- **Alibaba Trace Analysis**
  - Over-commitment in Alibaba datacenters
  - Average Utilization:
    - CPU: ~49%
    - Memory: ~82%
  - Static provisioning leads to under-utilization
  - CPU usage is strongly correlated with Memory
  - Container workloads show similar trend for CPU and Mem utilization
  - Correlation across utilization metrics is crucial for job placement

- **Kube-Knots: GPU-aware Orchestrator**
  - Kubernetes Head node
  - Knots GPU Monitor
  - Orchestrate Containers
  - Query every heartbeat
  - Container Resource Usage Profile
  - Docker Image Repository
  - We evaluate four GPU-specific scheduling schemes,
    - Uniform
    - Resource Agnostic Sharing
    - Correlation Based Provisioning
    - Peak Prediction

- **Analysis of GPU-aware cluster schedulers**
  - GPU load across three different workload mixes scheduled using Uniform scheduler
  - GPU load across three different workload mixes scheduled using Peak Prediction scheduler
  - Peak Prediction scheduler effectively consolidated the work-load mixes for all workloads

- **Overall Results and Discussion**
  - Peak Prediction (PP) improves the cluster-wide GPU utilization by up to 80% for both 50th and 90th percentile GPU utilization.
  - 33% savings in power consumption across the ten-node GPU cluster.
  - Reduced the overall QoS violations of latency sensitive queries by up to 53% when compared to the resource agnostic scheduler with GPU utilization.
  - PP scheduler’s misprediction rate (error) was as low as 16% for a window interval of 5s.

**ACKNOWLEDGMENTS**

This research was generously supported by several NSF grants 1320478, 1409095, 1629192, 1439021, 1629915, 1439567, 1626251, 1526750 and NSF Chameleon Cloud project CH-819640 for the GPU cluster infrastructure.

Prashanth Thinakaran (prashanth@cse.psu.edu)