On Power-Performance Characterization of Concurrent Throughput Kernels

Nilanjan Goswami, Yuhai Li, Amer Qouneh, Chao Li and Tao Li

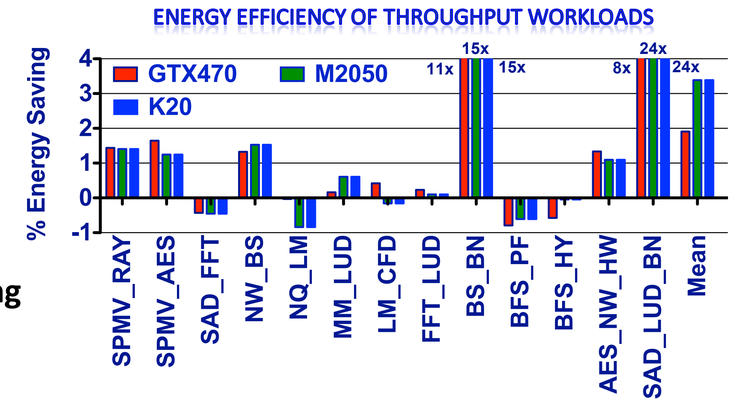
Intelligent Design of Efficient Architectures Laboratory (IDEAL), University of Florida

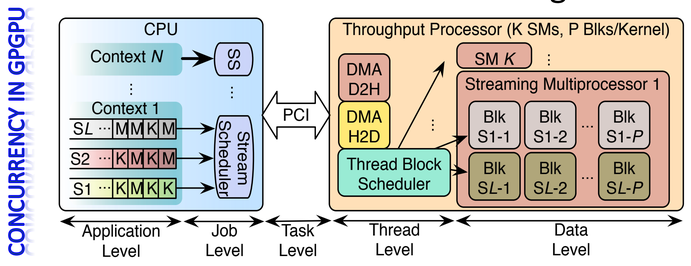
Introduction

- **GPUs are rapidly becoming datacenter backbone**
  - 90 supercomputers use GPUs (4 in Top10)
- **GPU the *Exascale Computing* "enabler"**
  - 10^{18} FLOPs require massively parallel processor
  - Enormous energy/Power budget of *exascale* requires energy/power efficient accelerators
  - Performance shouldn't be compromised
  - Synergy between power-performance is imperative
- **Perf/Power: Simultaneous kernels on GPU**
  - Concurrent kernels unlock power-performance co-optimization opportunity
  - Better resource utilization potentially can improve perf/power characteristics

Motivation

- **Energy efficient cloud based HPC data center**
  - Improved utilization drives
    - Power-performance co-optimization
    - Amortize long term operating cost
  - GPGPU workload exploration required
    - Improved concurrency implies better utilization
- **App, jobs and task concurrency analysis lacking**
  - Best combination of throughput kernels
  - Collective benefit vs. Individual gain

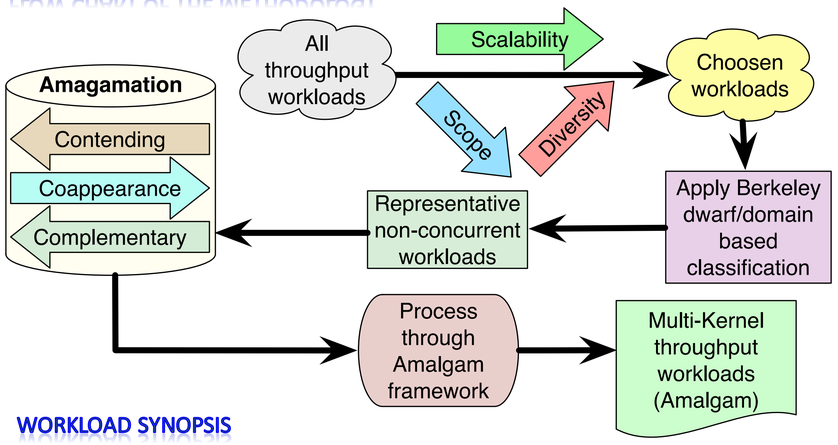




- **The concurrency within the GPU**
  - App, Job, Task, Thread and Data level
  - Various concurrency limiting factors in levels
  - Through investigation is necessary to know the factors and their contribution
  - Methodological exploration is missing too

Simultaneous Kernel Analysis

FLOW CHART OF THE METHODOLOGY



WORKLOAD SYNOPSIS

Benchmarks	Domain (Berkeley Dwarf)
Breadth First Search	General algorithms (Graph algorithm)
Sum of Absolute Difference	Video processing (Structured grid)
LU Decomposition	High performance computing (Dense linear algebra)
Matrix Multiplication	High performance computing (Sparse linear algebra)
Black Scholes	Finance (Monte Carlo method)
Binomial Options	Finance (Monte Carlo method)
Path Finder	General algorithms (Dynamic programming)
2D Convolution using FFT	Audio processing (Spectral method)
Ray Trace	Computer graphics (Graphical models)
Computational Fluid Dynamics Solver	High performance computing (Unstructured grid)
Sparse-Matrix Dense-Vector Multiplication	High performance computing/Image processing (Sparse linear algebra)
Heart Wall	Health application (Structured grid)
Hybrid Sort	General algorithms (Sorting)
Needleman-Wunsch	Health application (Dynamic programming)
N-Queen Solver	High performance computing (Back-track)
Advanced Encryption Standard	High performance computing/database (Combinational logic)
Lava MD2	High performance computing (Particle method)

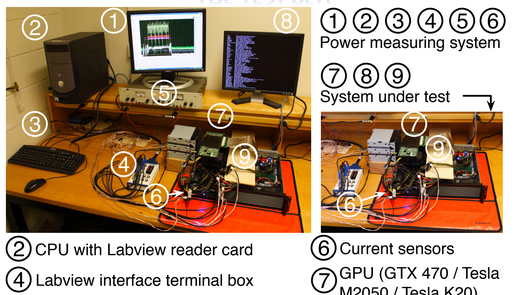
WORKLOAD INTRINSIC CHARACTERISTICS

Characteristics	Synopsys
Registers/Thread	Number of registers used per thread
Shared Memory	Amount of shared memory used per thread
Branch Efficiency	Percentage of non-divergent branches
Thread Batch Efficiency	Percentage of non-divergent thread batches
Kernel Count	Total number of kernels
Thread Count	Total number of threads launched
Dynamic Instructions	Dynamic instructions count across all kernels
Local Memory Inst.	Local memory load-store count
Global Memory Inst.	Global memory load-store count
Shared Memory Instruction	Shared memory load-store count
Branch Instructions	Total branch instructions count
Divergent Branches	Total divergent branch instructions count
Atomic Instructions	Total atomic instructions count
Device to Host Transfer	Device to host data transfer in bytes
Host to Device Transfer	Host to device data transfer in bytes
Off-chip Efficiency	Percentage off-chip row access locality

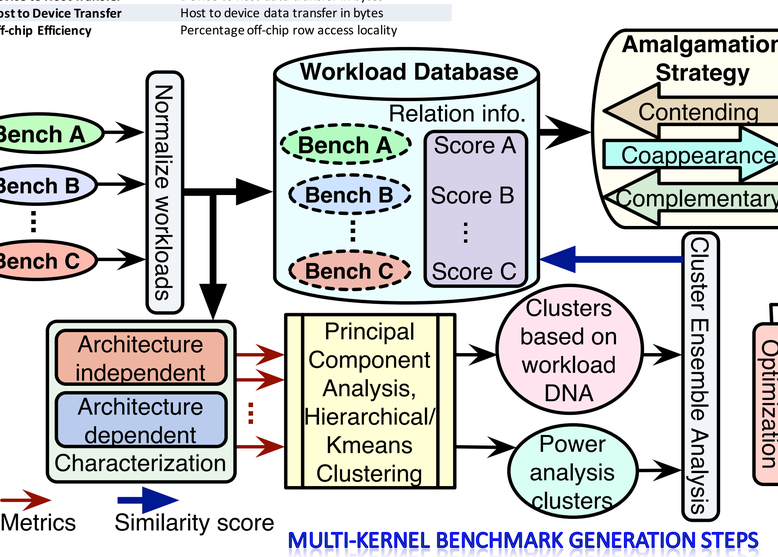
POWER-PERFORMANCE CHARACTERISTICS

Characteristics	Synopsys
Average Power	Average power across various kernels
Peak Power	Maximum power across various kernels
Total Energy	Total energy consumption for the workload
Instruction-per-Watt	Average power per instruction
Energy-Delay-Product	Energy multiplied by execution time
Instruction-per-Cycle	Average instructions executed per cycle
Instruction-per-Second	Average instructions executed per second
Execution Duration	Kernel execution time
Communication Overhead	Number of memory transfer commands
Maximum Temperature	Max temperature for fixed initial temperature

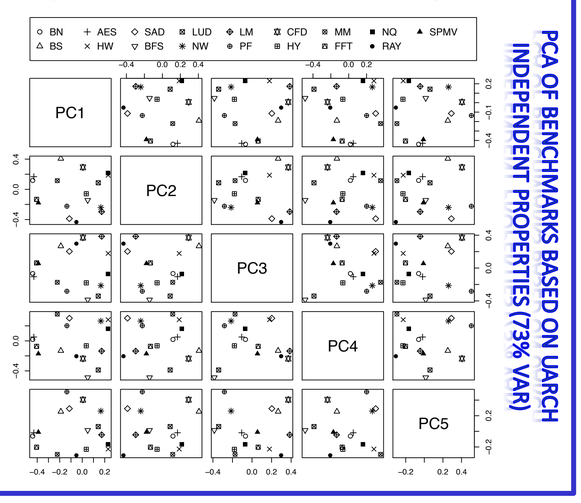
THE TEST BED



② CPU with Labview reader card
④ Labview interface terminal box
⑤ Bipolar power (+/- 12V DC) supply to sensor
⑥ Current sensors
⑦ GPU (GTX 470 / Tesla M2050 / Tesla K20)
⑨ Host CPU



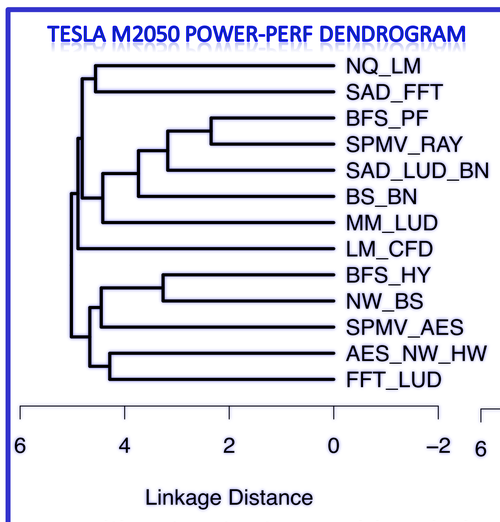
MULTI-KERNEL BENCHMARK GENERATION STEPS



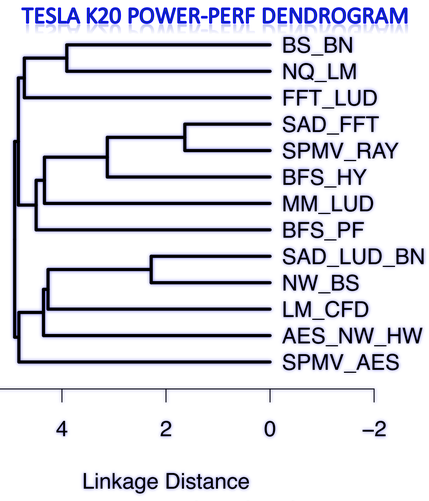
PCA OF BENCHMARKS BASED ON UARCH INDEPENDENT PROPERTIES (73% VAR)

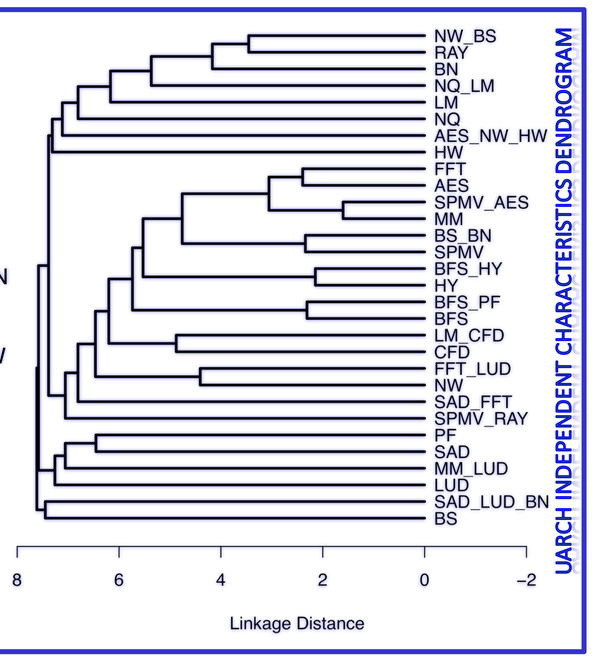
Evaluation

TESLA M2050 POWER-PERF DENDROGRAM



TESLA K20 POWER-PERF DENDROGRAM





- Unlike, intrinsic workload characteristics, individual workload dominance or subduing trend is absent in power-performance behavior of concurrent kernels

Conclusion

- **Presented a systematic multi-kernel GPGPU workload Perf/Power analysis method**
  - Based on performance, power, energy, utilization and interactions between them
  - Explored using real-world GPGPUs
- **Power profile and concurrency correlated**
  - Concurrency improves hardware utilization and helps in reducing energy
- **Diversity Analysis**
  - Using statistical analysis, demonstrated proposed workloads possess diversity

Future Work

- **Thorough study of the effects of concurrency on energy and power**
  - Feasibility of such concurrency and exploration of achievable overlap to improve energy efficiency
- **Power efficiency and occupancy analysis**

References

Pai, Sreepathi and Thazhuthaveetil, Matthew J. and Govindarajan, R., **Improving GPGPU Concurrency with Elastic Kernels**, in Proceedings of the Eighteenth International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS**), April 2013

IDEAL RESEARCH

NSF

HPC Center of the University of Florida

UF UNIVERSITY of FLORIDA
The Foundation for The Gator Nation