Argonne NATIONAL LABORATORY

MUMMI: EXPLORING PERFORMANCE AND POWER TRADEOFFS FOR PARALLEL APPLICATIONS

VALERIE TAYLOR, XINGFU WU
Mathematics and Computer Science Division
Argonne National Laboratory
The University of Chicago

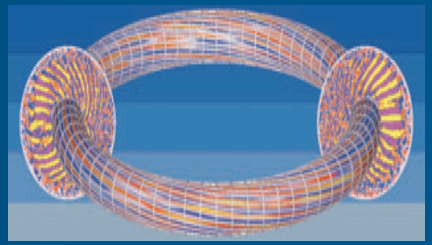
U.S. DEPARTMENT OF ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

2019 Scalable Tools Workshop
July 30, 2019, Lake Tahoe, CA

Energy Reduction

Scenario 1:

Power & Time



GTC, 100ppc
on SystemG

Methods: DVFS, DCT,
4x4 blocks

#Cores	GTC Type	Runtime (s)	Node Power (W)	Node Energy (KJ)
16x8 (128)	Hybrid	453	293.19	132.82
	Optimized-Hybrid	421 (-7.6%)	276.35 (-6.1%)	116.34 (-14.16%)
32x8 (256)	Hybrid	455	294.58	134.03
	Optimized-Hybrid	424 (-7.31%)	279.35 (-5.45%)	118.44 (-13.16%)
64x8 (512)	Hybrid	436	294.79	128.53
	Optimized-Hybrid	423 (-3.1%)	271.12 (-8.73%)	114.72 (-12.03%)

Energy Reduction

Scenario 2:

↑ Power, ↓ Time

MPI BT: Class B on SystemG
Method: Hybrid (MPI+OpenMP)

#Cores	BT Type	Runtime (s)	Node Power (W)	Node Energy (J)
4	MPI	269	281	58,643
	Hybrid	257 (-4.6%)	224.82 (3.13%)	57,779 (-1.47%)

Energy Reduction

Scenario 3:

↓ Power, ↑ Time

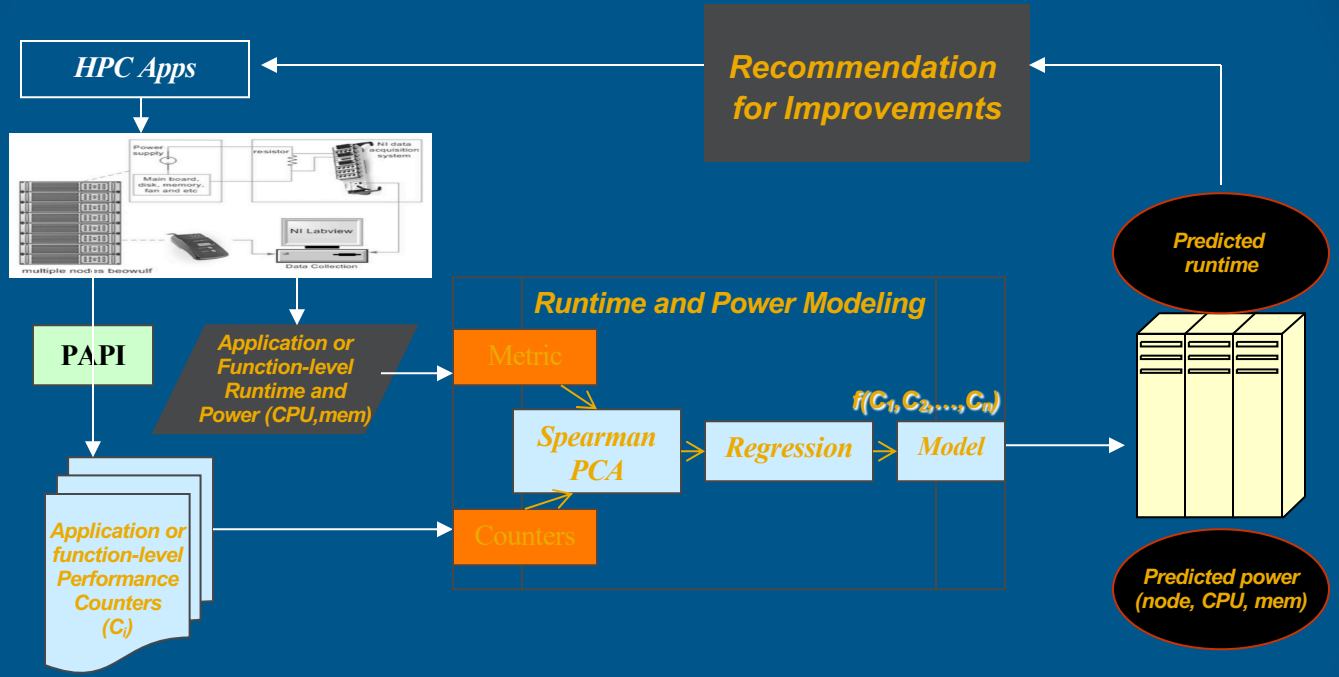
Hybrid BT: Class B on SystemG
Method: Frequency Change

#Cores	CPU Frequency	Runtime (s)	Node Power (W)	Node Energy (J)
16	1.8GHz	71.72	222.37	15,941.09
	1.2GHz	97.37 (35.76%)	148.34 (-33.29%)	14,444.04 (-9.39%)

Energy Saving Techniques

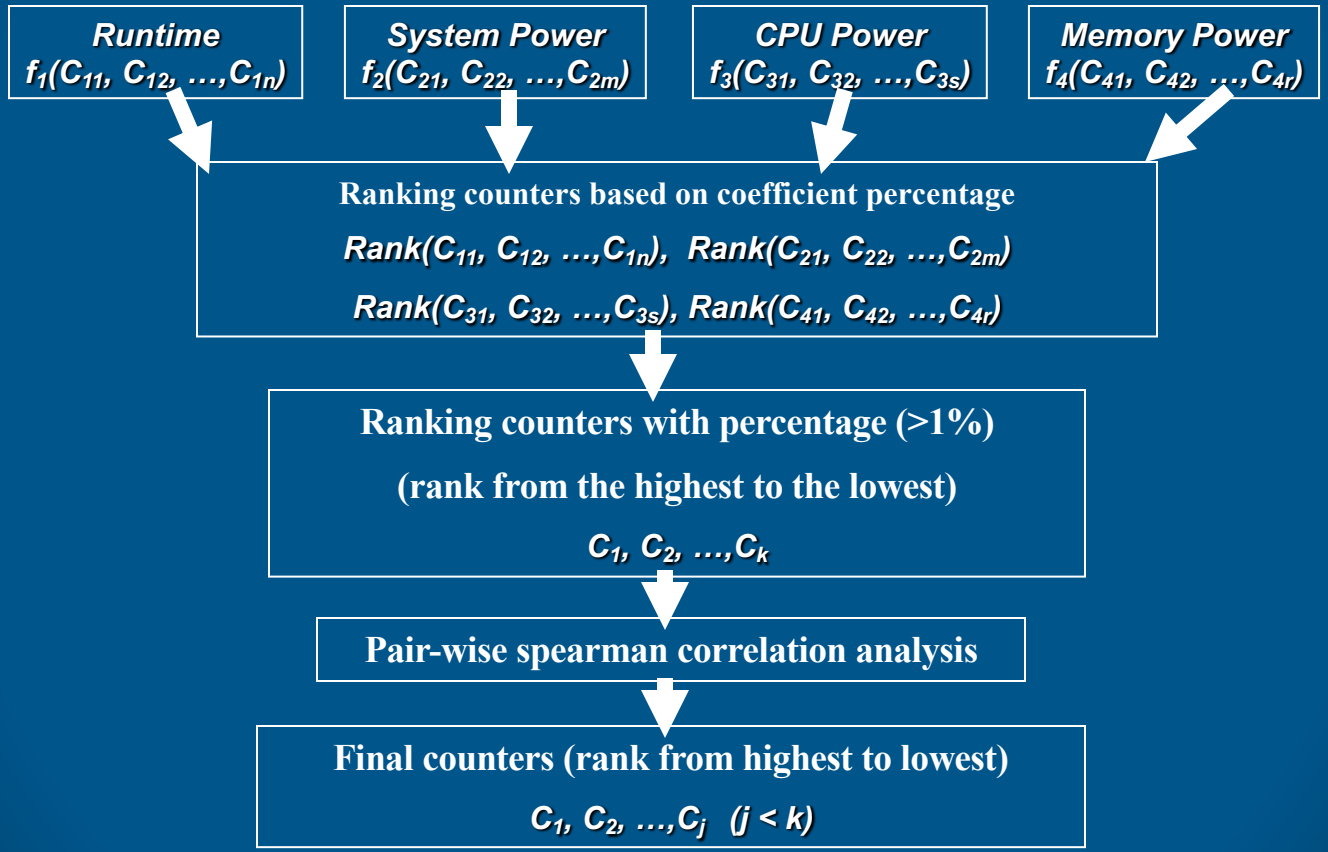
- **Hardware-based**
  - Energy efficient processor, memory, network, I/O
  - Dynamic resource sleeping techniques
- **Software-based**
  - Dynamic Voltage and Frequency Scaling (DVFS)
    [Lowenthal, Roundtree, Cameron]
  - Dynamic Concurrency Throttling (DCT)
    [de Supinski, Shultz, Olivier, Roundtree]
- **Application-based**
  - Power capping
  - Utilize hardware counters for hints

MuMMI: Performance Counter-based Modeling



Four metrics: runtime, node power, CPU power, memory power

Counter-Guided Application Refinements

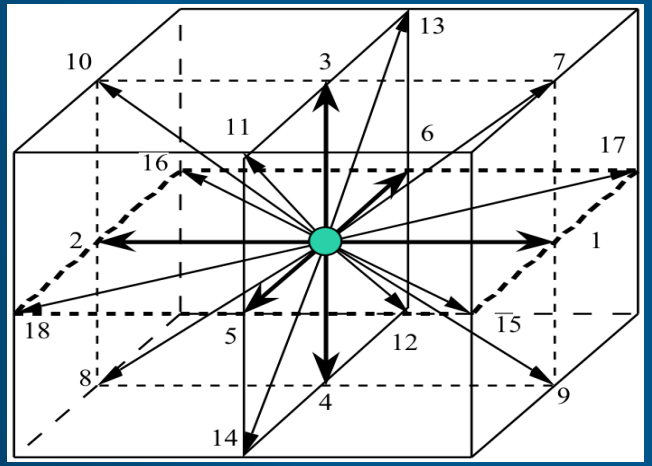


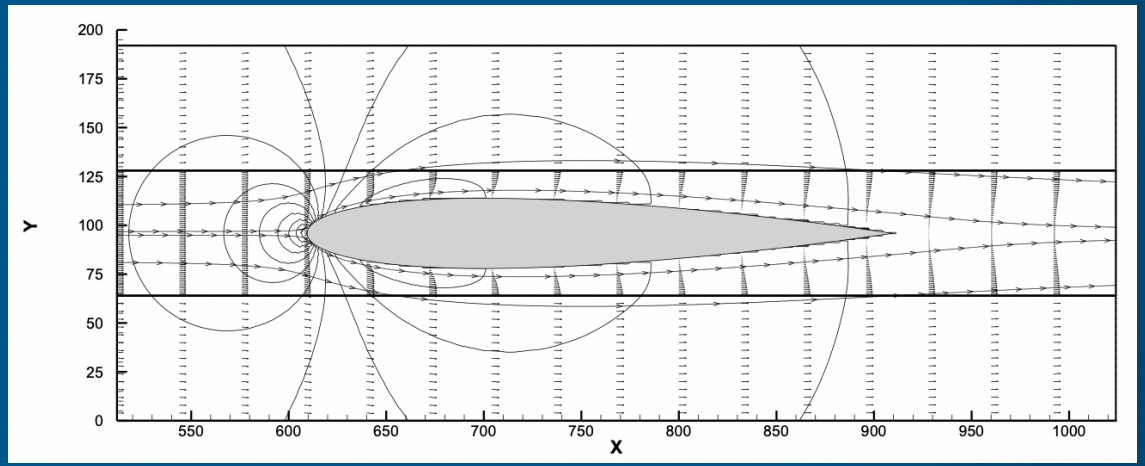
Power-aware Parallel Systems

	Mira	SystemG	Theta
Time Frame	2013- present	2008-2015	2017- present
Total Nodes	49,152	324	4,392
Cores/Node	16	8	64
CPU Type	IBM Power BQC 16C 1.6GHz	Intel Xeon 2.8GHz Quad-Core	Intel Xeon Phi KNL 7230 1.30GHz
Memory/Node	16GB	8GB	192GB DDR/ 16GB MCDRAM
L1 Inst./Data Cache per core	16KB/16KB	32KB/32KB	32KB/32KB
L2 Cache/Chip	32MB	12MB	32MB
Interconnect	5D Torus	QDR Infiniband	Cray Aries Dragonfly
Location	Argonne	Virginia Tech	Argonne

PMLB: Parallel Multi-block Lattice Boltzmann

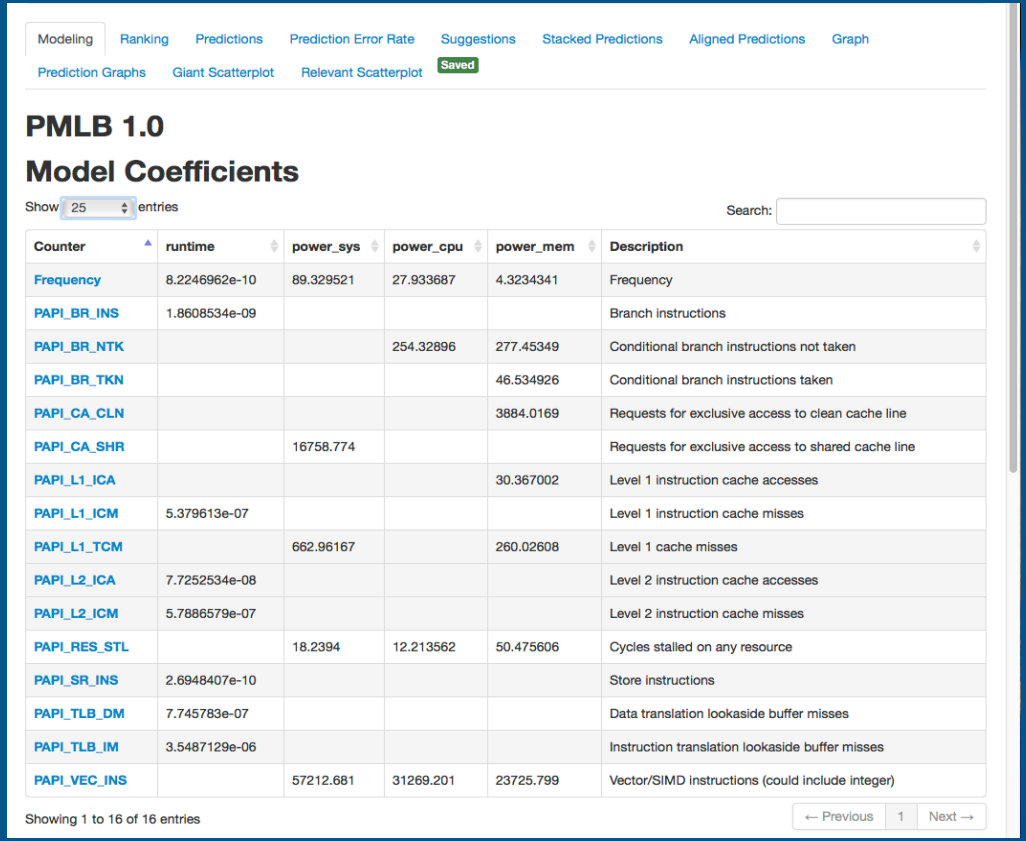
- Lattice Boltzmann Method (LBM): widely used for fluid dynamics
- Aerospace application uses Q19D3 lattice model to simulate uniform flow over airfoil
  - 19 velocities in 3D, with the collision and streaming operations





PMLB: 128x128x128 on SystemG

Runtime	System Power	CPU Power	Memory Power
$f_1(C_{11}, C_{12}, \ldots, C_{1n})$	$f_2(C_{21}, C_{22}, \ldots, C_{2m})$	$f_3(C_{31}, C_{32}, \ldots, C_{3s})$	$f_4(C_{41}, C_{42}, \ldots, C_{4r})$



Modeling | Ranking | Predictions | Prediction Error Rate | Suggestions | Stacked Predictions | Aligned Predictions | Graph

Prediction Graphs | Giant Scatterplot | Relevant Scatterplot | Saved

PMLB 1.0

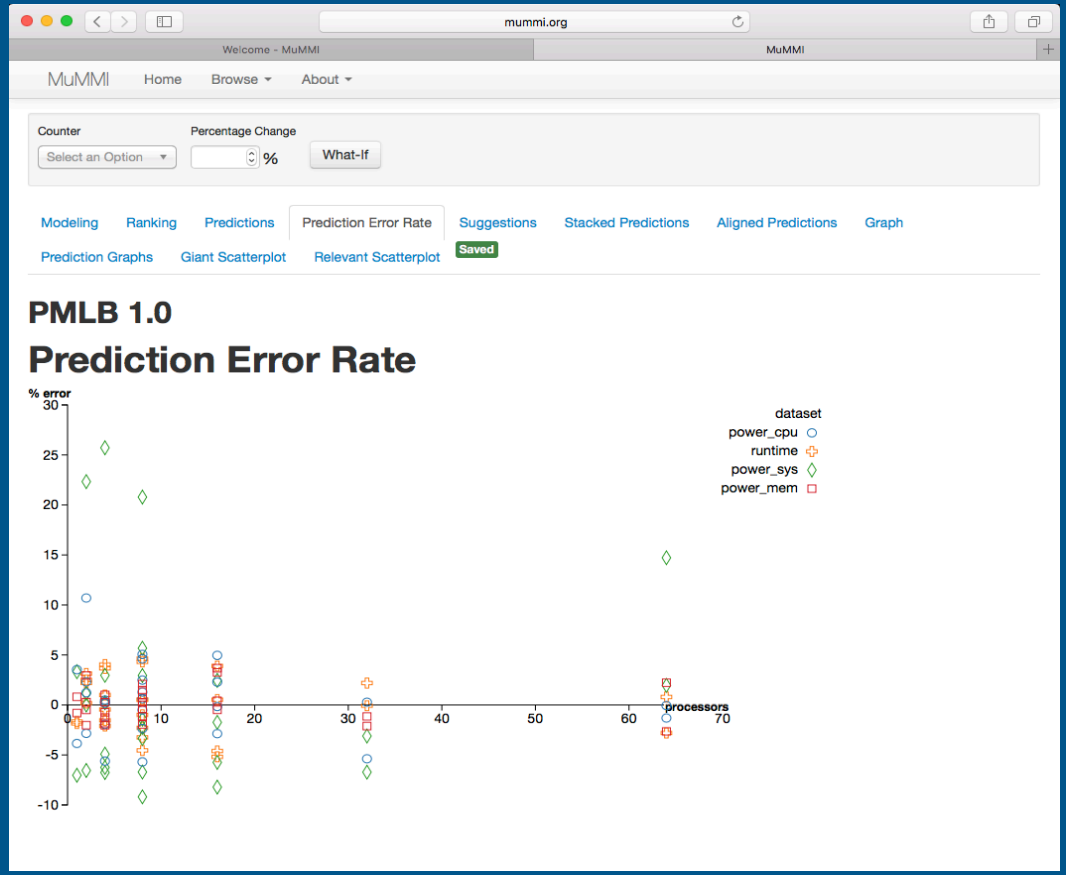
Model Coefficients

Show 25 entries Search:

Counter	runtime	power_sys	power_cpu	power_mem	Description
Frequency	8.2246962e-10	89.329521	27.933687	4.3234341	Frequency
PAPI_BR_INS	1.8608534e-09				Branch instructions
PAPI_BR_NTK			254.32896	277.45349	Conditional branch instructions not taken
PAPI_BR_TKN				46.534926	Conditional branch instructions taken
PAPI_CA_CLN				3884.0169	Requests for exclusive access to clean cache line
PAPI_CA_SHR		16758.774			Requests for exclusive access to shared cache line
PAPI_L1_ICA				30.367002	Level 1 instruction cache accesses
PAPI_L1_ICM	5.379613e-07				Level 1 instruction cache misses
PAPI_L1_TCM		662.96167		260.02608	Level 1 cache misses
PAPI_L2_ICA	7.7252534e-08				Level 2 instruction cache accesses
PAPI_L2_ICM	5.7886579e-07				Level 2 instruction cache misses
PAPI_RES_STL		18.2394	12.213562	50.475606	Cycles stalled on any resource
PAPI_SR_INS	2.6948407e-10				Store instructions
PAPI_TLB_DM	7.745783e-07				Data translation lookaside buffer misses
PAPI_TLB_IM	3.5487129e-06				Instruction translation lookaside buffer misses
PAPI_VEC_INS		57212.681	31269.201	23725.799	Vector/SIMD instructions (could include integer)

Showing 1 to 16 of 16 entries ← Previous | 1 | Next →

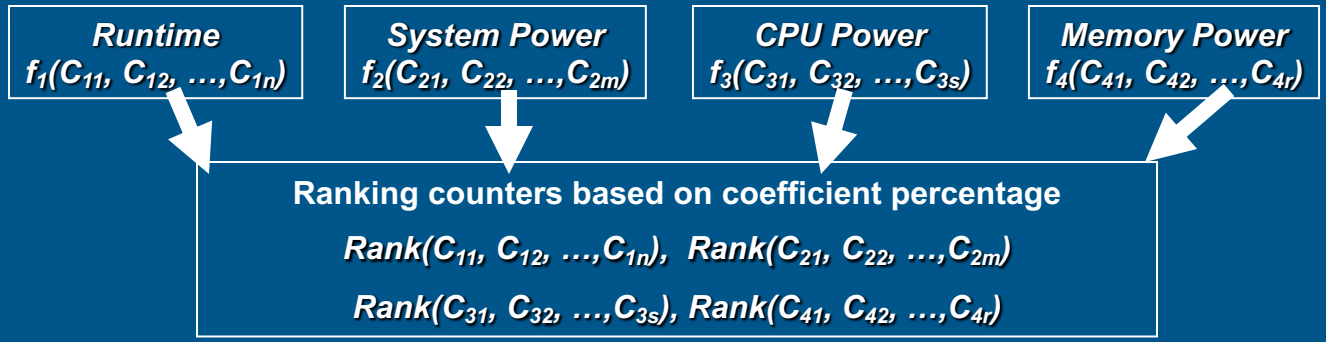
PMLB on SystemG: Prediction Error



www.mummi.org

Argonne NATIONAL LABORATORY

Counter Ranking



For example, given PMLB:

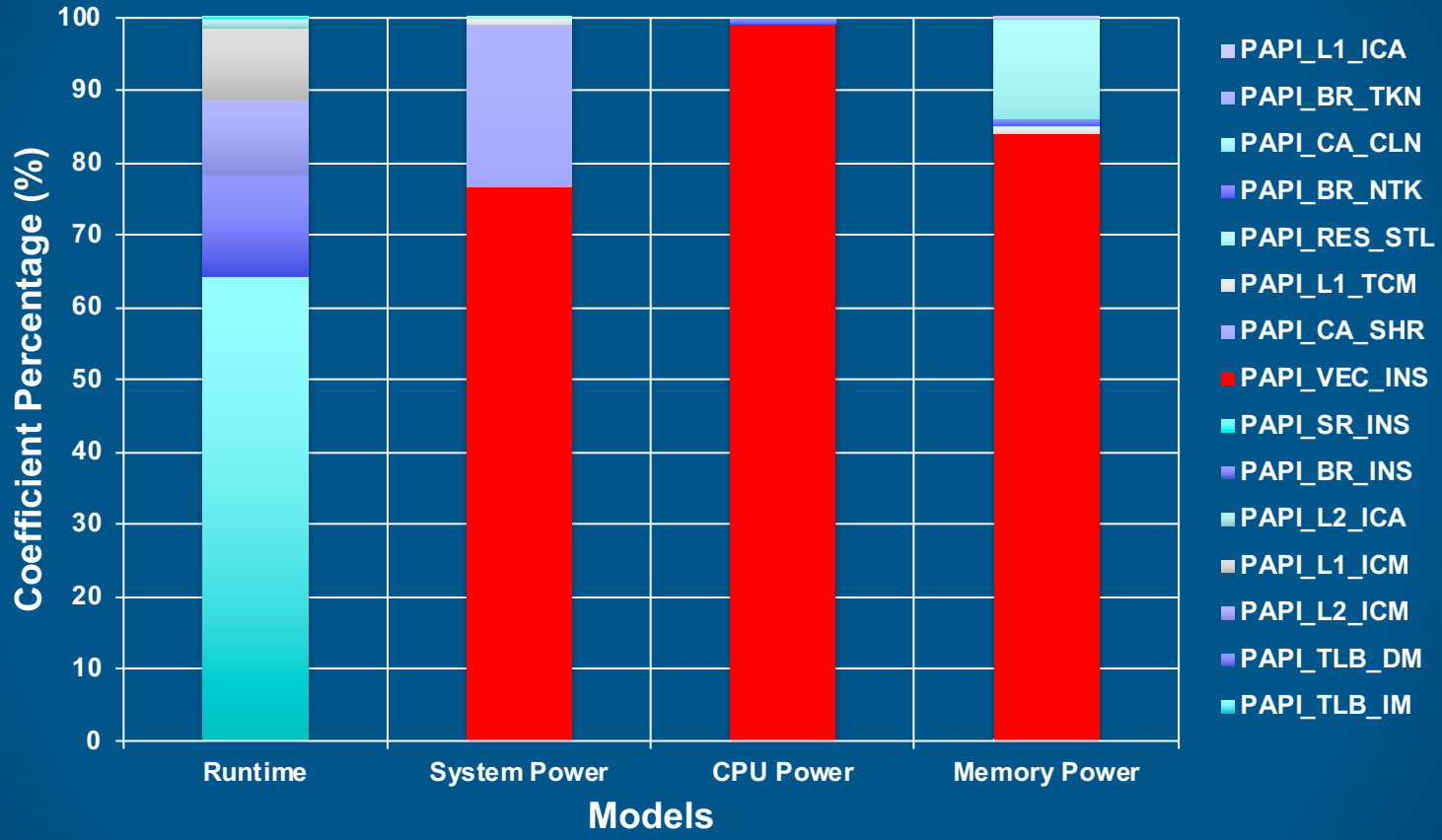
Runtime
TLB_IM: 64.29%
TLB_DM: 14.03%
L2_ICM: 10.49%
L1_ICM: 9.75%
L2_ICA: 1.40%
BR_INS: 0.03%
SR_INS: 0.01%

Node Power
VEC_INS: 76.64%
CA_SHR: 22.45%
L1_TCM: 0.89%
RES_STL: 0.02%

CPU Power
VEC_INS: 99.15%
BR_NTK: 0.81%
RES_STL: 0.04%

Memory Power
VEC_INS: 83.91%
CA_CLN: 13.74%
BR_NTK: 0.98%
L1_TCM: 0.92%
RES_STL: 0.18%
BR_TKN: 0.16%
L1_ICA: 0.11%

Counter Ranking for Original PMLB on SystemG



Correlation Analysis

- **TLB_IM: Occurred in P**
  - TLB_DM: Corr Va
  - BR_NTK: Corr V
  - L2_ICM: Corr
  - L1_ICM: Corr V
  - L2_ICA: Corr V
  - BR_TKN: Corr Valu
  - BR_INS: Corr Value=0.

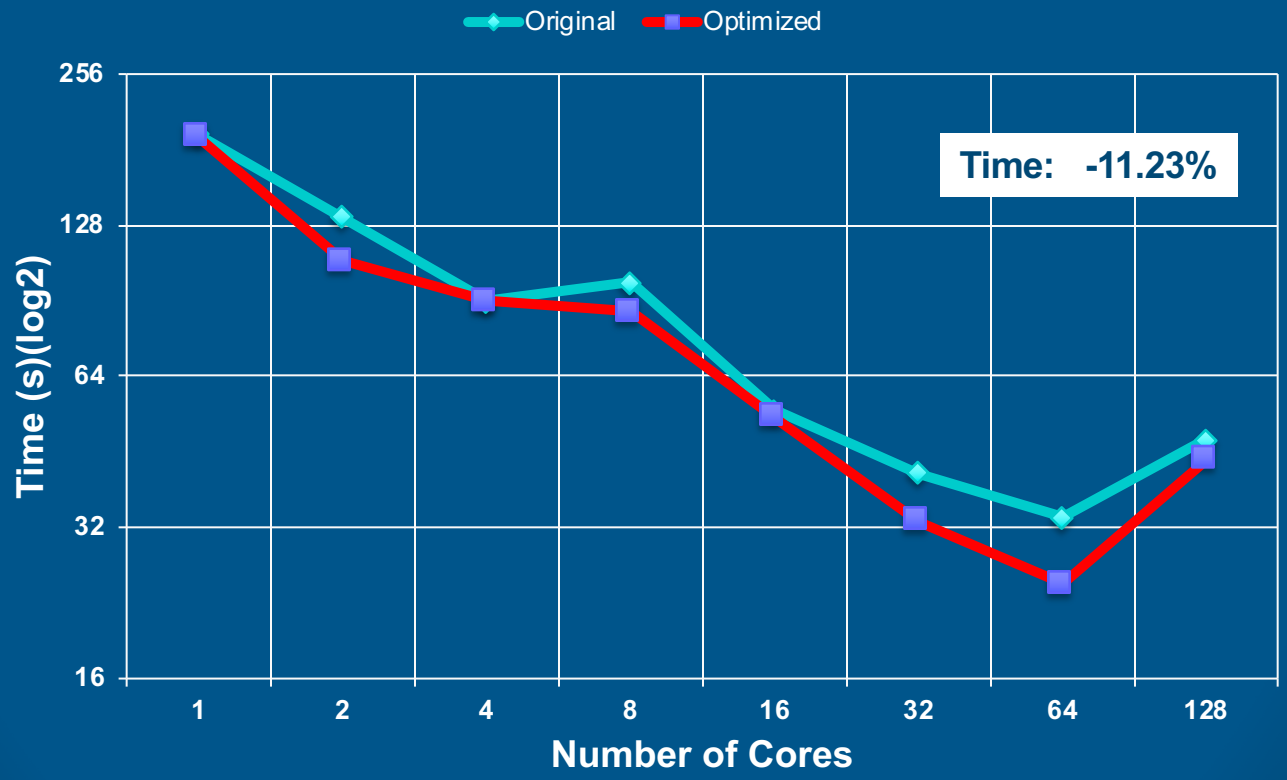
Final counters: TLB_IM and VEC_INS for optimization focus

- **VEC_INS: Occurred in System, CPU, Memory**

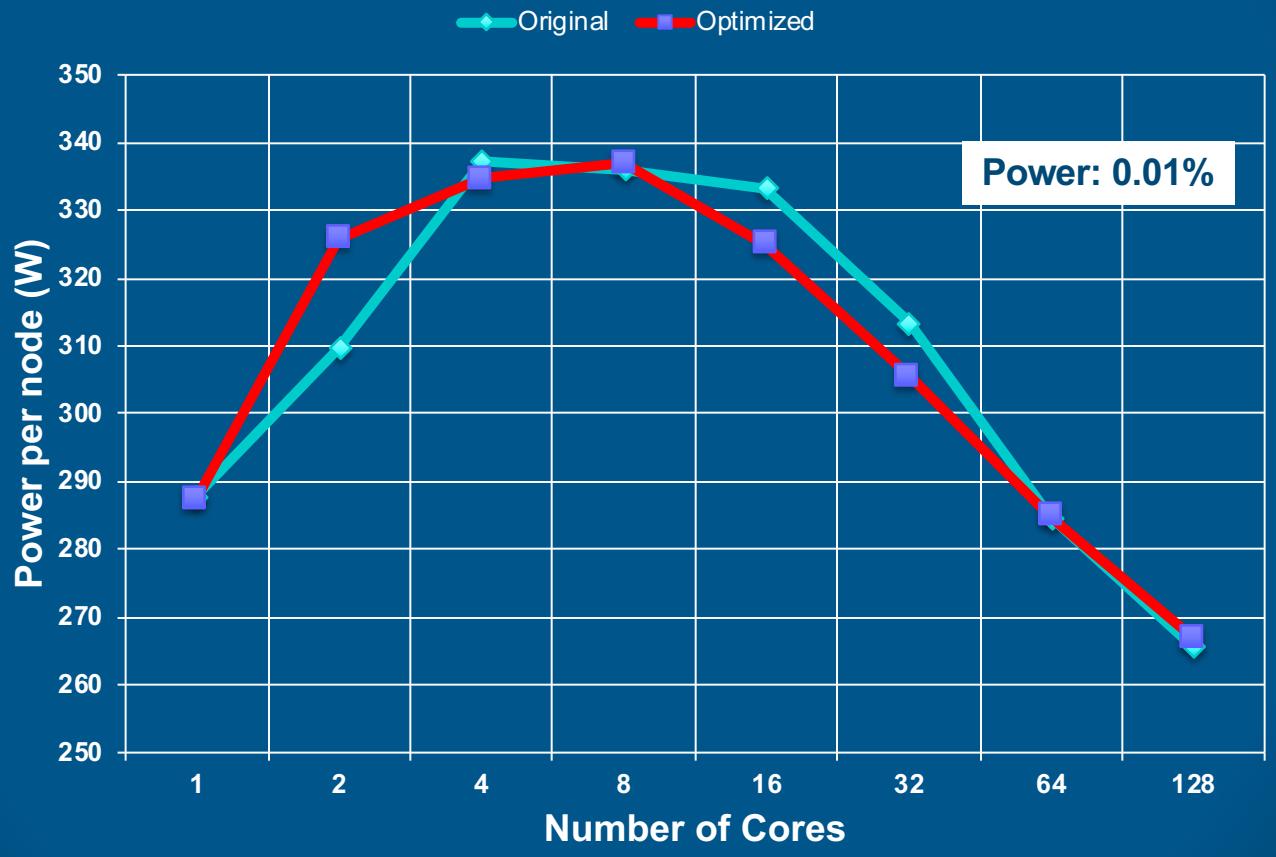
Counters-Guided Optimizations on SystemG

- TLB_IM (TLB Instruction Misses)
  - Use the 2MB huge pages for the application execution through the use of libhugetlbfs to reduce TLB misses
- VEC_INS (Vector Instructions)
  - Vectorize dominant code sections
    - streaming: 5 major loops
  - Use the compiler option *-ftree-loop-distribution*
    - perform loop distribution to improve cache performance and further vectorization

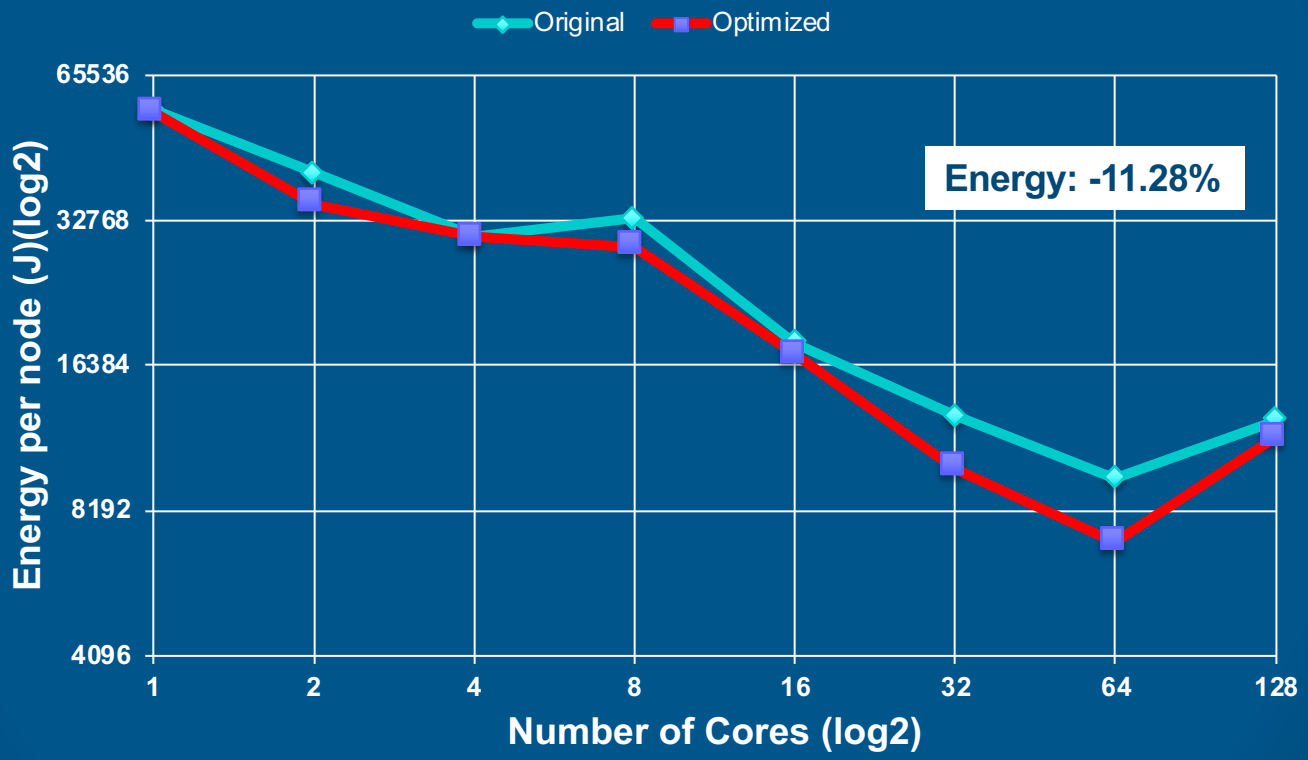
Performance for PMLB with 128x128x128 on SystemG



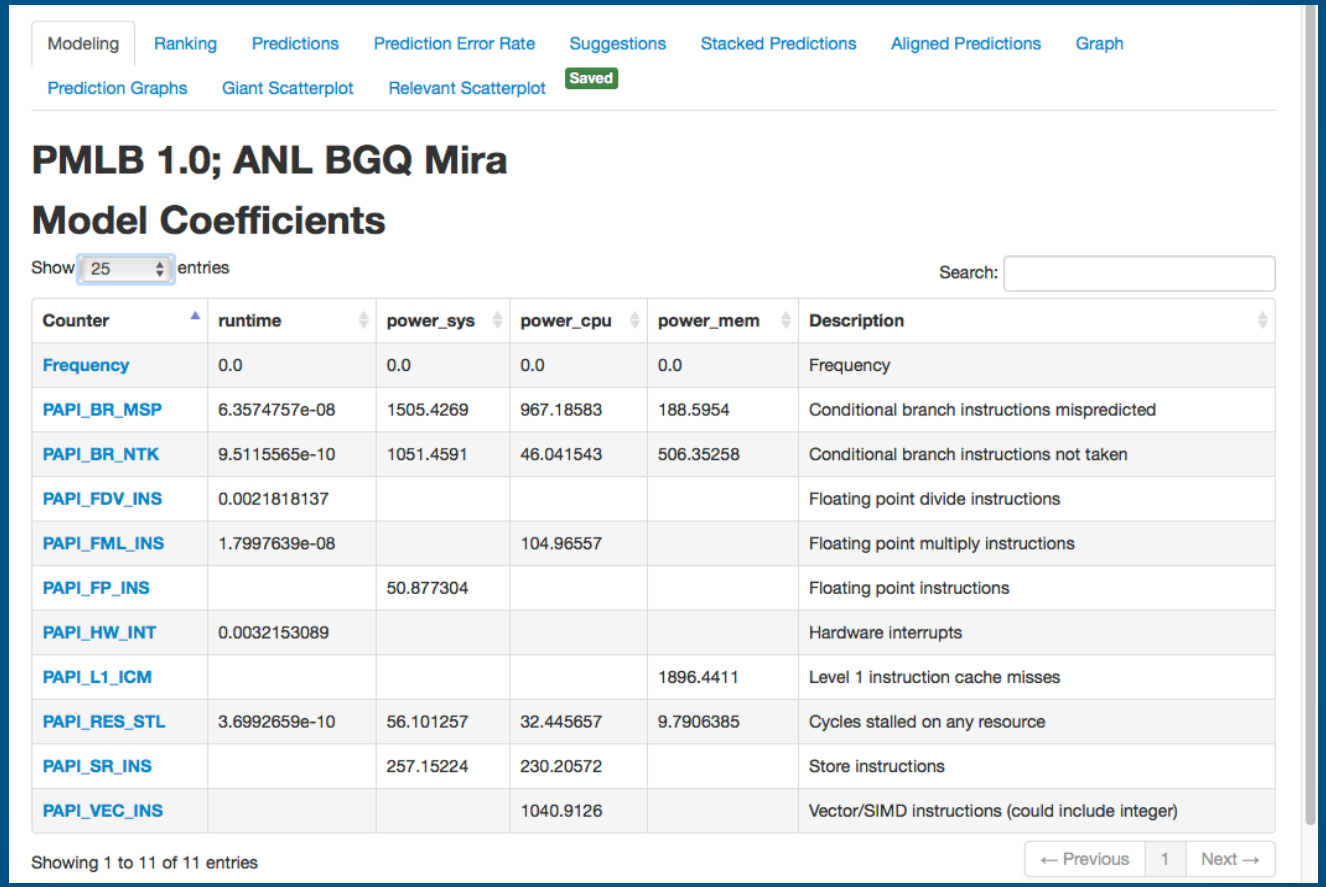
Node Power Comparison on SystemG



Energy Comparison for PMLB on SystemG



PMLB: 512x512x512 on Mira



Modeling | Ranking | Predictions | Prediction Error Rate | Suggestions | Stacked Predictions | Aligned Predictions | Graph

Prediction Graphs | Giant Scatterplot | Relevant Scatterplot | Saved

PMLB 1.0; ANL BGQ Mira

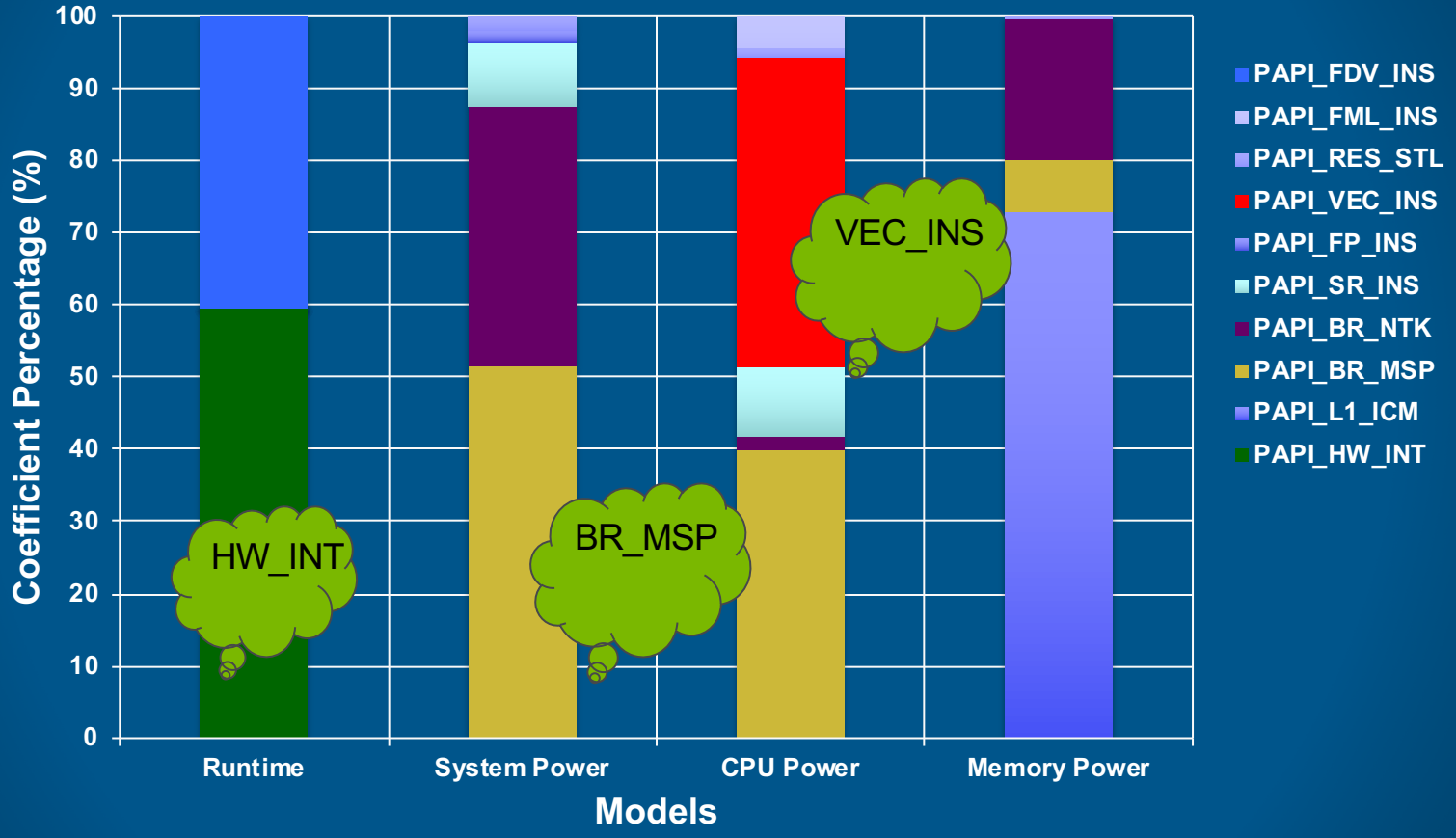
Model Coefficients

Show 25 entries Search:

Counter	runtime	power_sys	power_cpu	power_mem	Description
Frequency	0.0	0.0	0.0	0.0	Frequency
PAPI_BR_MSP	6.3574757e-08	1505.4269	967.18583	188.5954	Conditional branch instructions mispredicted
PAPI_BR_NTK	9.5115565e-10	1051.4591	46.041543	506.35258	Conditional branch instructions not taken
PAPI_FDV_INS	0.0021818137				Floating point divide instructions
PAPI_FML_INS	1.7997639e-08		104.96557		Floating point multiply instructions
PAPI_FP_INS		50.877304			Floating point instructions
PAPI_HW_INT	0.0032153089				Hardware interrupts
PAPI_L1_ICM				1896.4411	Level 1 instruction cache misses
PAPI_RES_STL	3.6992659e-10	56.101257	32.445657	9.7906385	Cycles stalled on any resource
PAPI_SR_INS		257.15224	230.20572		Store instructions
PAPI_VEC_INS			1040.9126		Vector/SIMD instructions (could include integer)

Showing 1 to 11 of 11 entries ← Previous | 1 | Next →

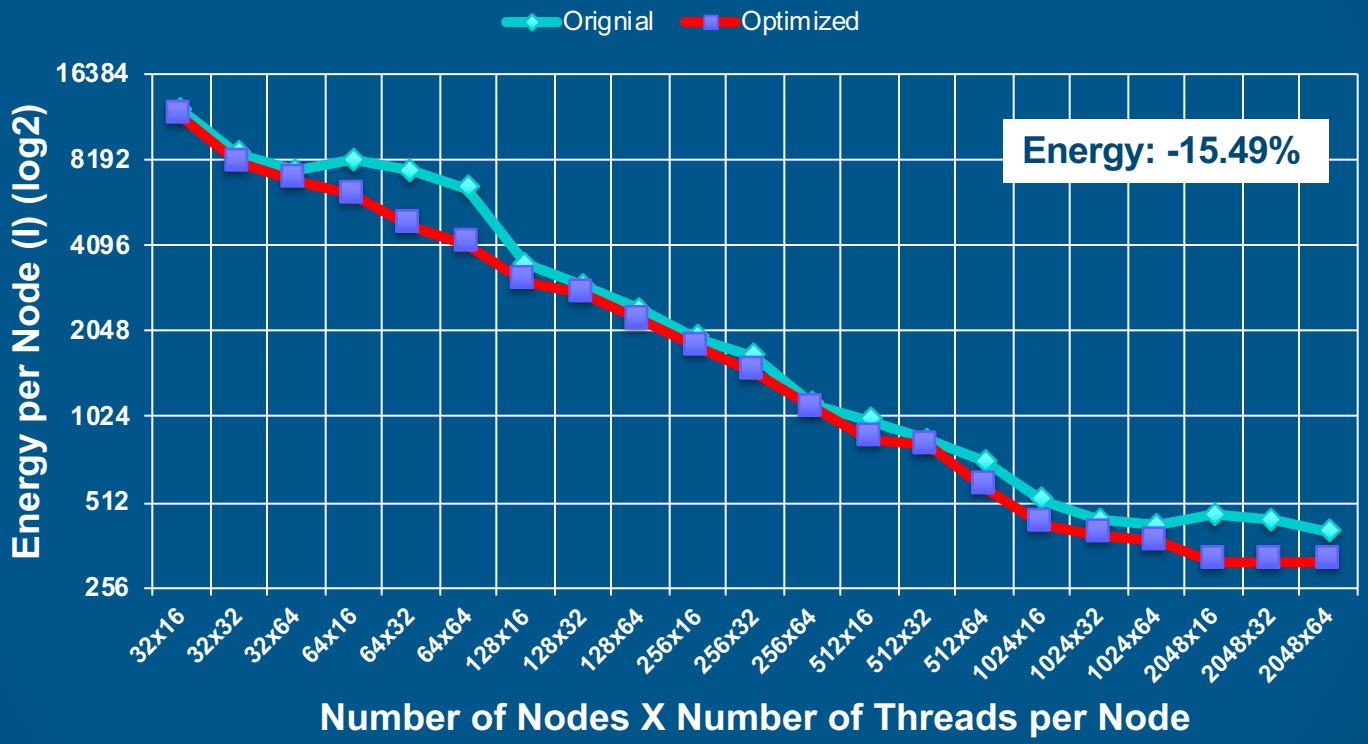
Counter Ranking for Original PMLB on Mira



Counters-Guided Optimizations on Mira

- HW_INT (Hardware Interrupt)
  - Inline several procedures
- BR_MSP (Branch Mispredictions)
  - Unroll several loops and eliminate some conditional branches
- VEC_INS (Vector Instructions)
  - Vectorize dominant code sections
    - streaming: 5 major loops
  - Use the compiler option –qarch=qp –qsimd=auto
    - Utilize the quad FPU to accelerate vector operations

Energy Comparison for PMLB with 512x512x512 on Mira



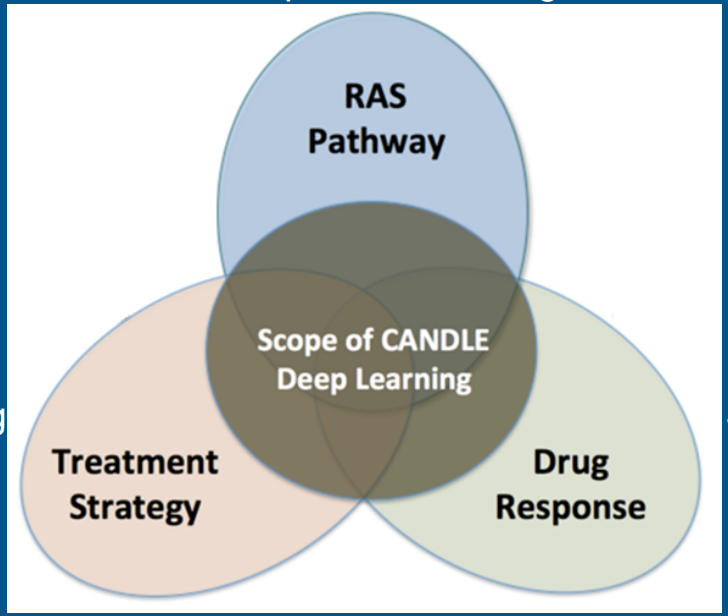
Additional Applications

App	System	# Cores	Dominant Counters	Energy Savings	Time	Power
eq3dyna	Mira	4,096	VEC_INS, BR_MSP	61.73%	↓	↑
eq3dyna	SystemG	256	L1_ICM, L2_ICA	30.67%	↓	↑
Lulesh	Shepard	864	L2_ICM, L3_TCM	58.30%	↓	↓
BT-MZ	SystemG	512	Cache_FLD,TOT_INS	14.71%	↓	↓
SP-MZ	SystemG	512	L2_TCH, TOT_INS	16.75%	↓	↓

Current Work: ML Application

CANDLE (Cancer Distributed Learning Environment) focuses on building a single scalable deep neural network code that can address three cancer challenge problems:

Unsupervised learning



Semi-supervised learning

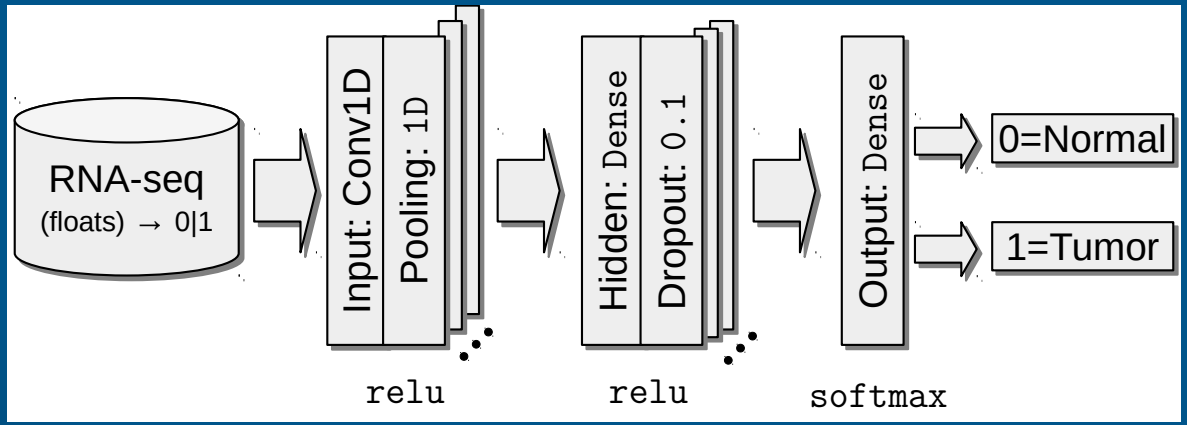
Supervised learning

Source: CANDLE, http://candle.cels.anl.gov

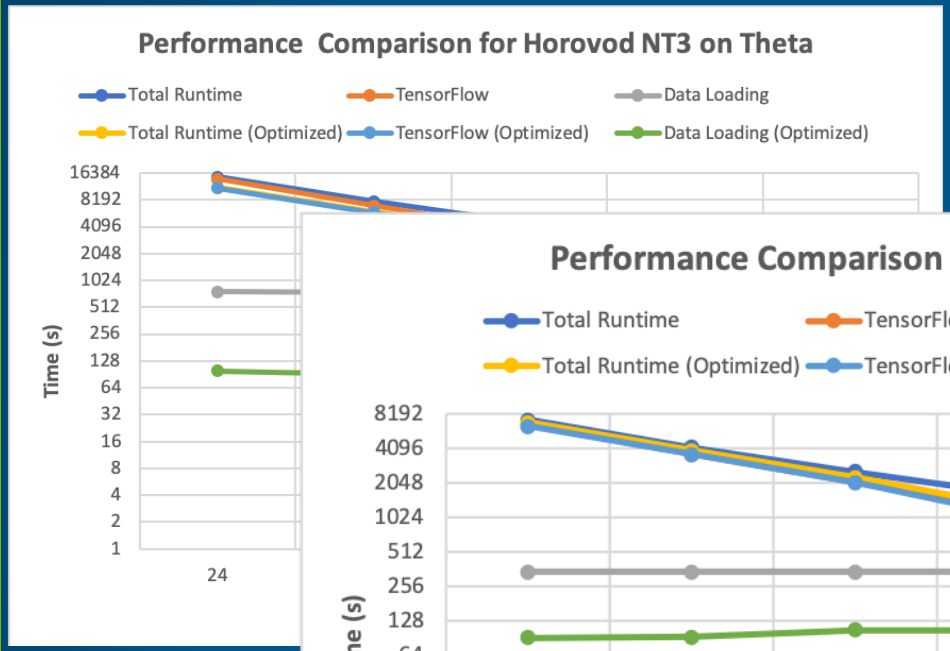
CANDLE Benchmarks

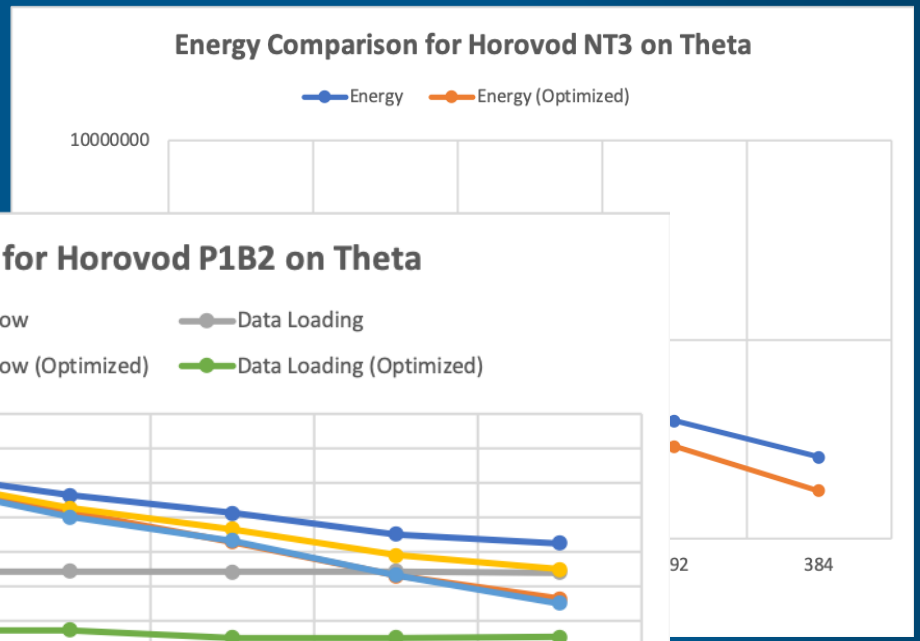
- **Pilot 1 Benchmarks**: P1B1, **P1B2**, P1B3, **NT3** (TensorFlow)
  - At the cellular level to predict drug response based on molecular features of tumor cells and drug descriptors
  - Implementation: Python, TensorFlow, Keras; Library: MKL-DNN, cuDNN

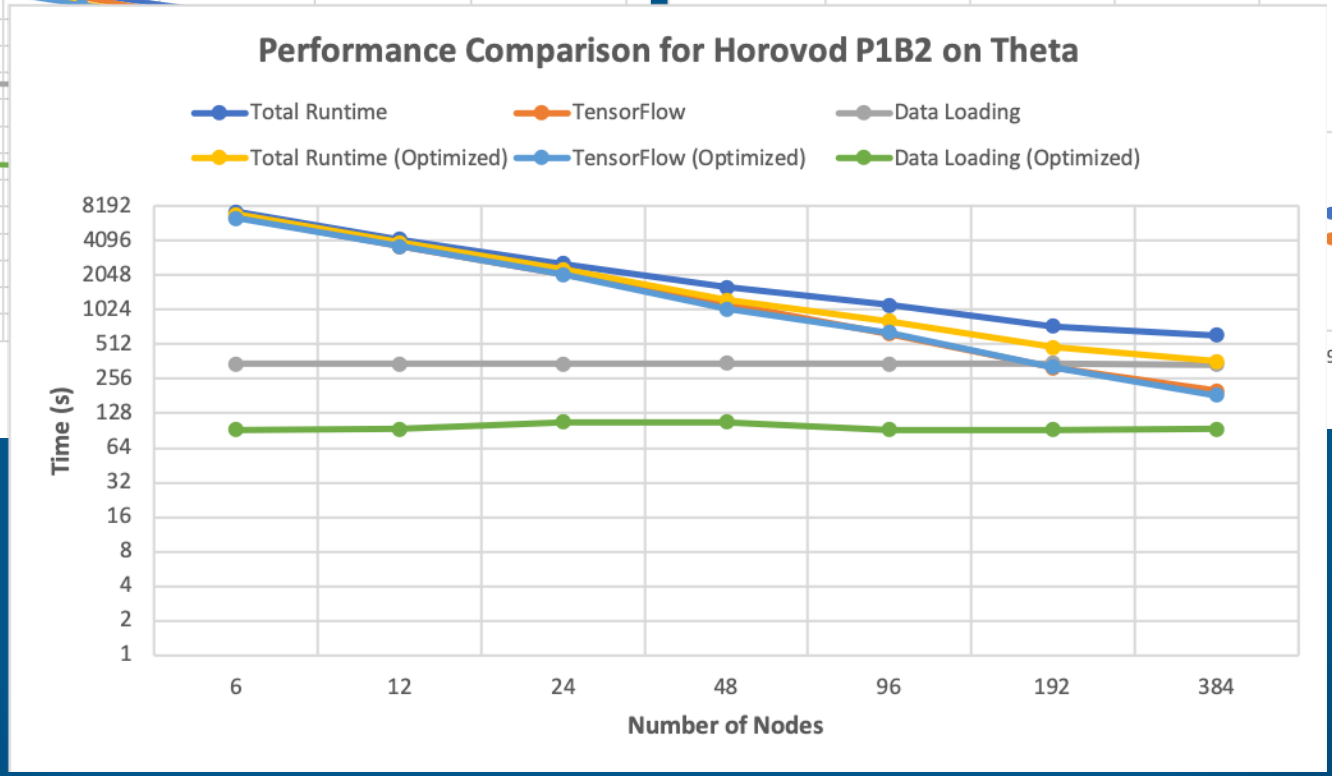
Architecture of NT3 Benchmark



Source: CANDLE Benchmarks, https://github.com/ECP-CANDLE/Benchmarks







Applying MuMMI to P1B2 for More Improvement

www.mummi.org/eamom/238

MuMMI Home Browse About

Counter: Select an Option | Percentage Change: % | What-If

Modeling | Ranking | Predictions | Prediction Error Rate | Suggestions | Stacked Predictions | Aligned Predictions | Graph | Prediction G

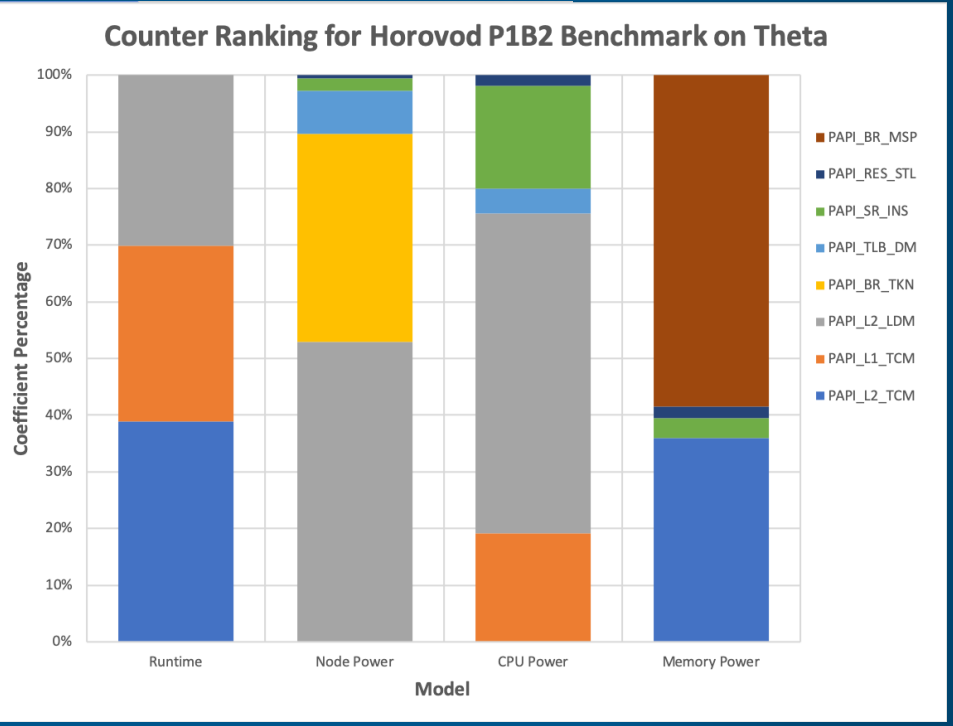
Horovod CANDLE P1B2 Benchmark 1.0

Model Coefficients

Show 10 entries

Counter	runtime	power_sys	power_cpu	power_mem
Frequency	0.0	0.0	0.0	0.0
PAPI_BR_MSP				492.03848
PAPI_BR_TKN		3064.9412		
PAPI_L1_TCM	1.1886415e-07		711.91732	
PAPI_L2_LDM	1.1531975e-07	4420.747	2091.2069	
PAPI_L2_TCM	1.4923964e-07			302.42454
PAPI_RES_STL		43.600267	68.719892	17.171306
PAPI_SR_INS		182.1489	672.47233	30.428916
PAPI_TLB_DM		641.28111	160.67402	
PAPI_TOT_INS		29.533281		

Showing 1 to 10 of 10 entries



Summary

- Application Refinement for Energy Reduction
  - Explore performance counter-based models for runtime and power requirements and tradeoffs
  - Identify the most important counters for application refinement
  - Refine the applications for better energy efficiency

- Further Work
  - Use performance counter-based modeling to further improve the performance and energy of caner deep learning applications
  - Explore behavior of different classes of applications (more cancer deep learning applications) on different architectures such as ORNL Summit.
  - Explore power capping to improve energy efficiency of applications