Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Building `Cooler` MVAPICH2

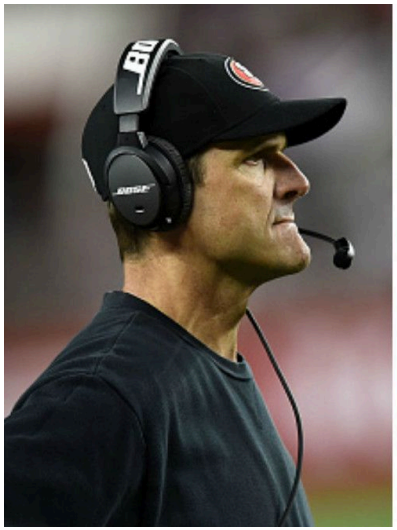
ABHINAV VISHNU

MUG' 15

August 19-21, 2015

U.S. DEPARTMENT OF
ENERGY

Buckeyes: The Undisputed National Champions



November 28th, 2015

Acknowledgement

Akshay Venkatesh

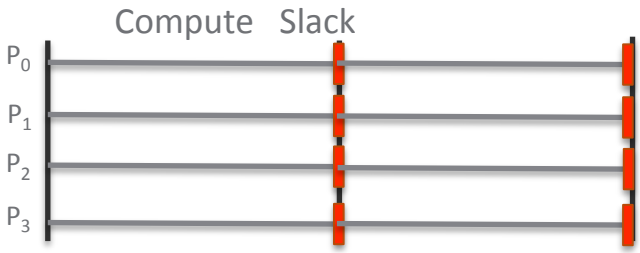
MUG'13

Summer 2014

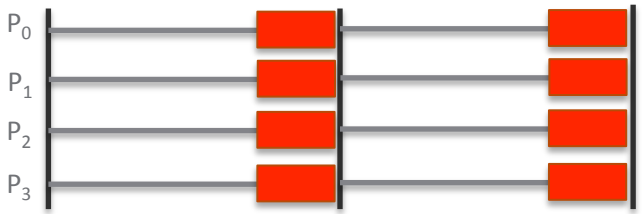
MUG'15,
SC'15 – Best Student Paper Finalist

Contributors:
Khaled Hamidouche
Nathan Tallent
DK Panda
Darren Kerbyson
Adolfy Hoisie

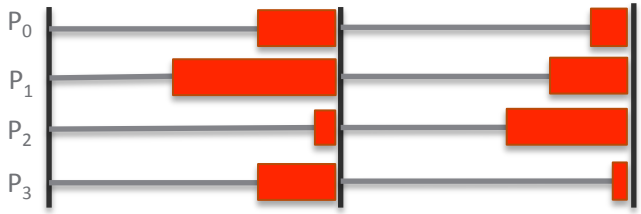
Motivation: Application Patterns



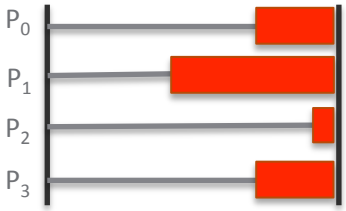
An application writer's wish



What he thinks is happening,
Iterative-Temporal Pattern



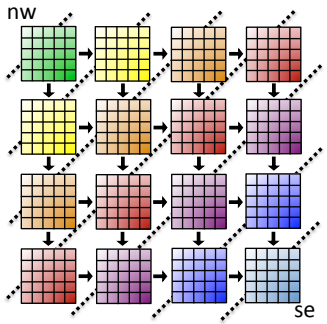
What is actually happening,
Iterative Non-Temporal Pattern



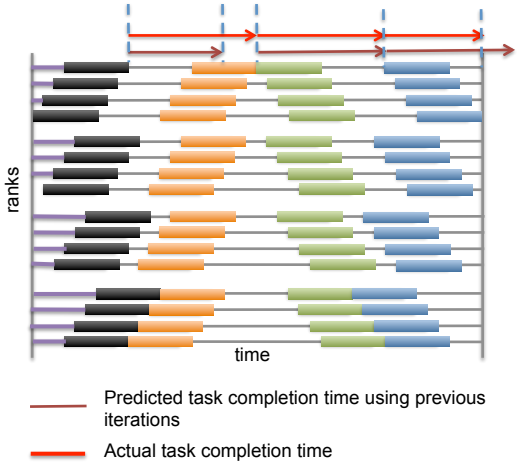
Non-iterative application

Existing Literature and Limitations

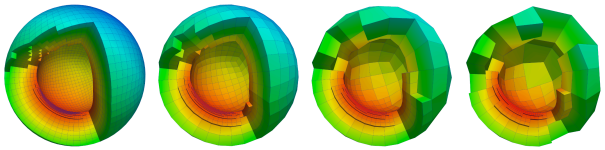
- Adagio
  - LLNL runtime for saving energy
  - Slows down tasks (period between MPI calls) using DVFS
  - Fails for several patterns
- Existing MVAPICH(2)
  - Collective algorithms for large data transfer
  - Most applications use small data
- Per-call methods
  - Sundriyal et al.
- Using Historical information
  - PMAC's Green Queue

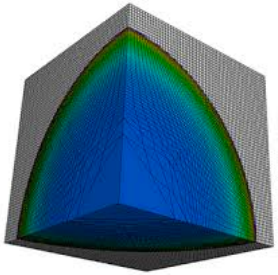


(a)



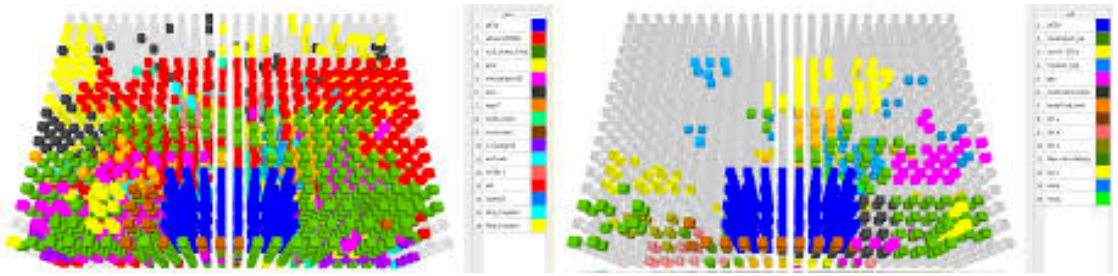
(b)





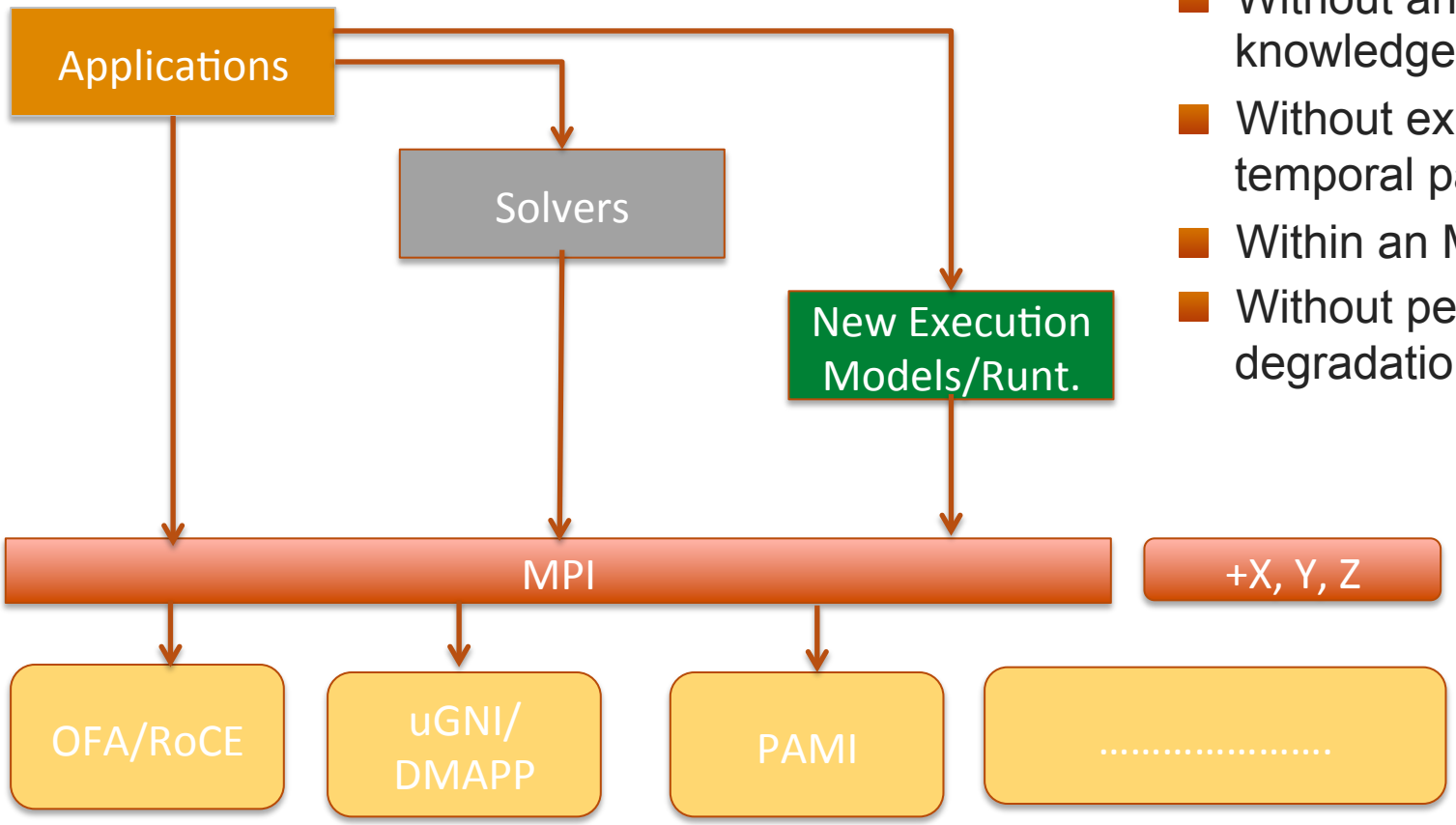
Let's say we could model these applications …

- System factors
  - OS noise
    - SC'03 (Petrini et al.)
    - SC'10 (Hoefler et al.)
  - Network Congestion
    - PPoPP'15 (Tallent et al.)
    - CCPE'10 (Vishnu et al.)
    - SC'13 (Bhatele et al.)
  - System faults
    - IPDPS'15 (Song et al.)
- Combinations of system factors
- ***Static modeling is insufficient!***



Bhatele et al. – SC'13

MPI is the Lowest Common Denominator

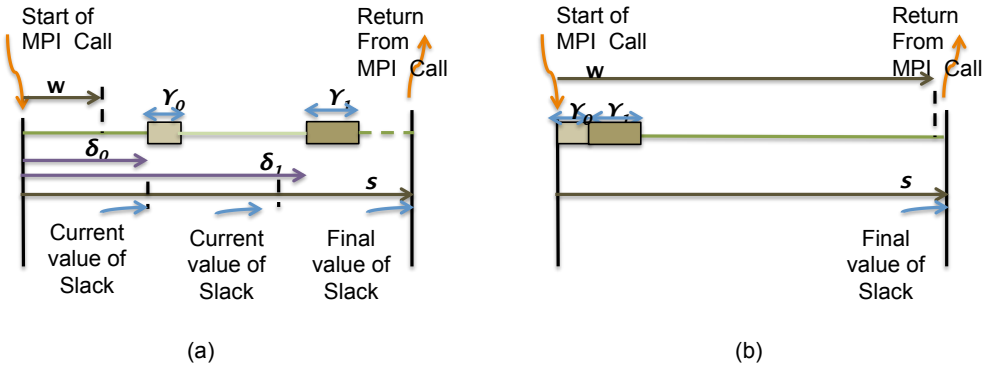


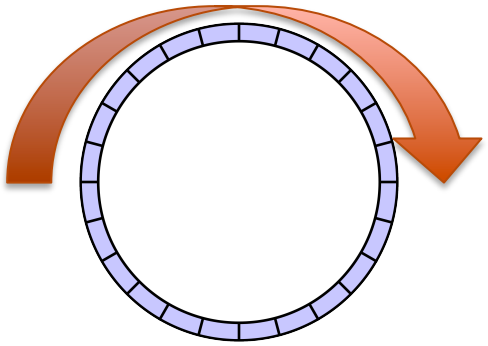
- Can we save the energy
  - Without any application knowledge,
  - Without expecting temporal patterns,
  - Within an MPI runtime,
  - Without performance degradation?

Communication Protocols from Energy Standpoint

- Eager-Send
  - A false positive
  - Copy and return
- Receive
  - Unexpected Queue search
  - Channel search
    - Network and Shared Memory
  - Use Power Lever
- Non-blocking messages
  - False Positives
- Progress Primitives
  - Consider a circular buffer of requests
  - Use time-out to continue search

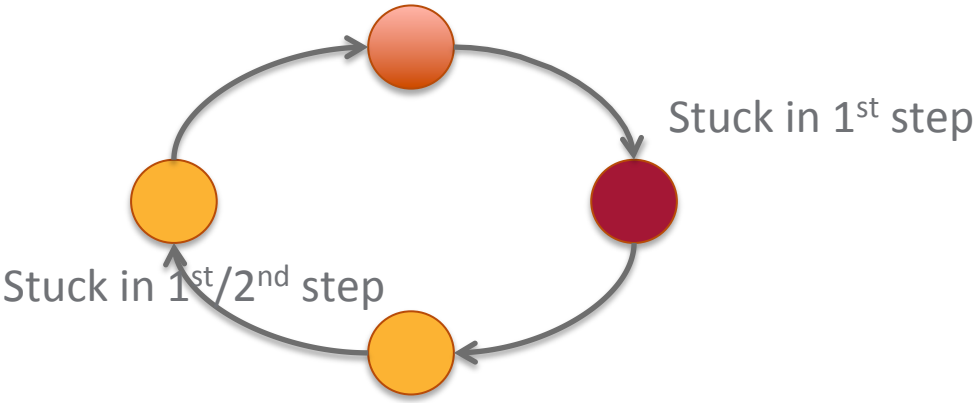
An Example of using LogGP for maximizing True positives in EAM



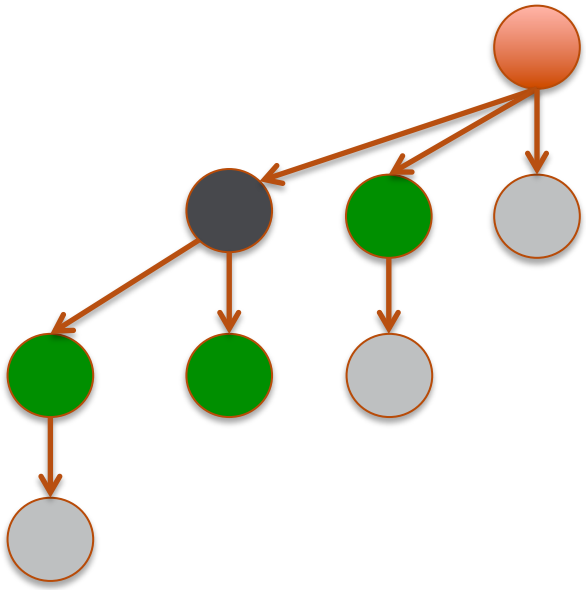


Energy Efficient Collective Communication

Un-rooted Collectives



Rooted Collectives

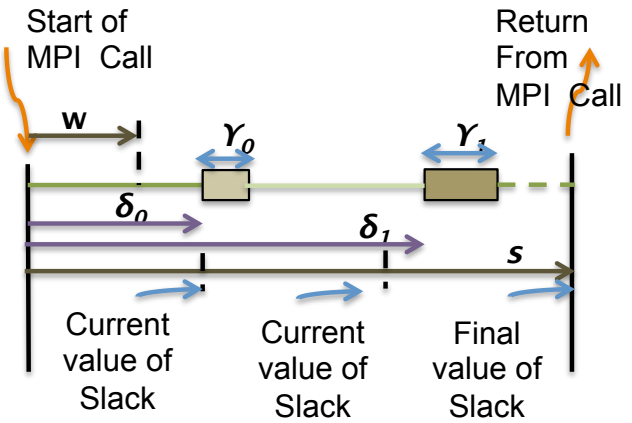


A process can observe slack in 1st or 2nd step, but none after that!

Processes can use communication modeling to use power levers at start Of an MPI call

Bounding Performance and Energy Loss

- Bounding performance loss
  - A power lever is used *only after* its threshold is crossed
  - Worst case occurs, when completion occurs right after a power lever is applied
  - Performance never degrades beyond user-specified threshold
  - In many cases, the cost of applying a lever is further amortized
- Bounding loss of energy savings
  - Worst case is when completion occurs right before a threshold can be breached



(a)

Performance Evaluation

- Experimental Testbed
  - TACC Stampede
  - 6400 Nodes of Intel Sandybridge (Xeon E5-2680) + 1 MIC co-processor
  - Two level Fat Tree, InfiniBand FDR
- MVAPICH2 – 2.0.2
- Design Points
  - Optimistic (Current Default), Pessimistic and Energy Aware MPI (EAM)
- Power Levers
  - Interrupt driven execution (5us overhead, 66% power improvement)
  - ~~DVFS (Dynamic Voltage and Frequency Scaling)~~
    - Not available on any production systems (TACC, PNNL Cascade)
- Power Measurement
  - Modified Intel Running Average Power Limit (RAPL)
    - Distributed collection of on-board power counters
    - Handle wrap-around with time-out based collection

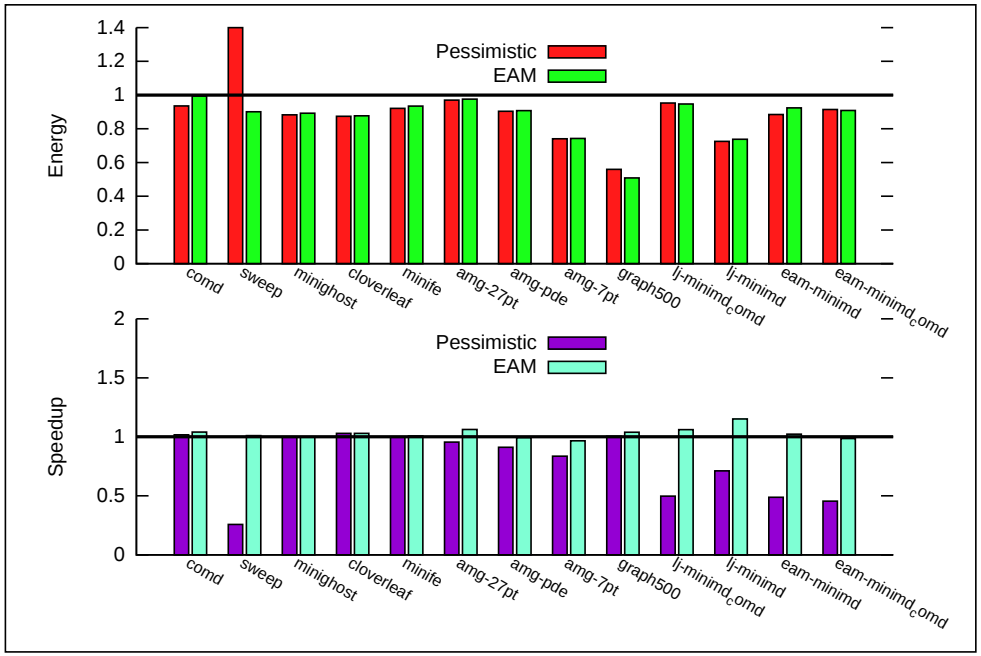
Summary of Results

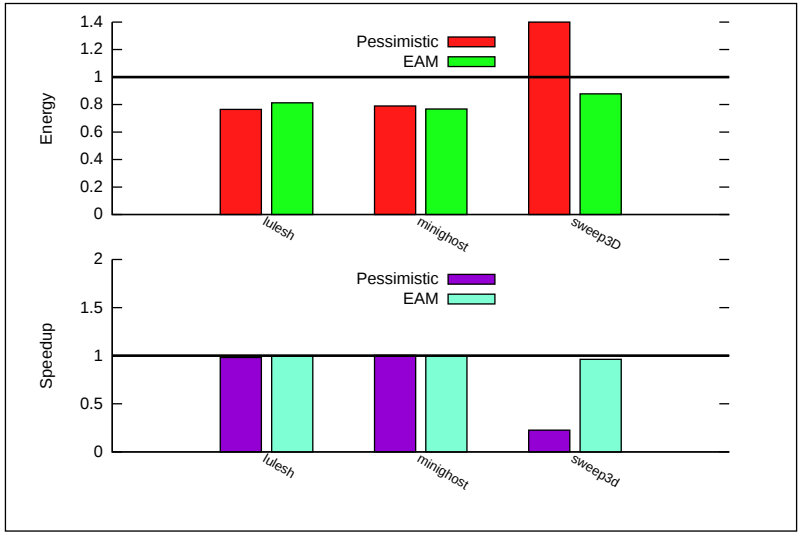
	Application	Objective	Major MPI calls	Class	Benefits(%)	Loss(%)
1	miniFE	Unstructured	Allreduce	Iterative/Temporal	25%	1%
2	miniMD	Molecular dynamics	Create_cart, Barrier	Iterative/Temporal	26%	2%
3	miniGhost	FDM/FVM	Waitany, Allreduce	Iterative/Temporal	23%	0%
4	CloverLeaf	Euler equations on grids	Allreduce	Iterative/Temporal	12%	3%
5	CoMD	Molecular Dynamics	SendRecv, Barrier	Iterative/Temporal*	4%	4%
6	Hoomd-Blue	Many-paricle dynamics	Allreduce, Bcast	Iterative/Temporal	5%	0%
7	AMG	Parallel AMG	Allreduce, Allgather	Iterative/Non-Temporal	10%	1.15%
8	Sweep3D	Parallel Neutron Transport	Recv	Iterative/Non-Temporal	12%	1%
9	LULESH	hydrodynamic equations	Allreduce	Iterative/Non-Temporal	18%	0.5%
10	Graph500	Breadth-first search	Alltoall	Non-Iterative	41%	4%

*Application thought to be iterative-temporal, and behaves as non-temporal

- We can save energy on 10 different applications --- without knowing their internal computation and communication behaviour
  - And that is a good thing!
- Performance loss is <= 4%
- Several applications are in 12-25% energy savings range
  - Without sweating for it

Evaluation at 2K and 4K processes





- EAM is able to save similar energy as Pessimistic design
  - In several cases, Pessimistic degrades performance, while EAM does not
- Performance benefits on three class of applications indicates that EAM will likely be beneficial for other applications!

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Abhinav Vishnu

abhinav.vishnu@pnnl.gov

http://hpc.pnl.gov/people/vishnu