Improving MVAPICH to Enable Science to Improve Our World

6th Annual MVAPICH User Group Meeting

Adam Moody
Livermore Computing

August 7, 2018

LLNL-PRES-756121

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

The `M' in MUG

MVAPICH source code based on MPICH and MVICH

MPICH

MPI (on) Chameleon

Message Passing Interface on Chameleon

~~MVICH~~

~~MPI for Virtual Interface Architecture (on Chameleon)~~

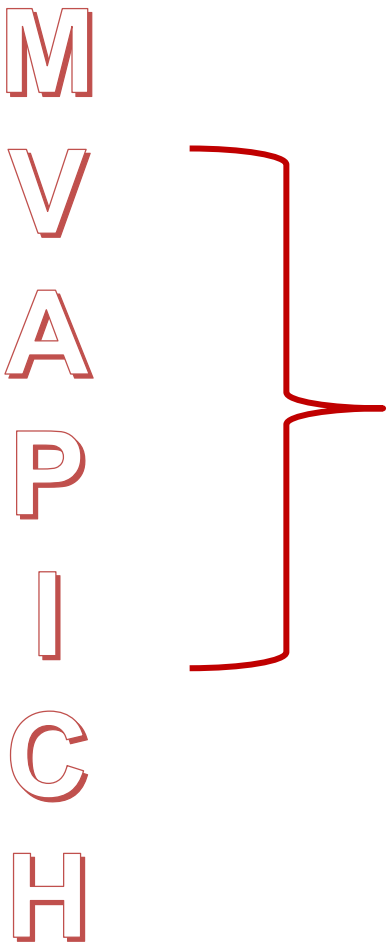
an MPI implementation over the InfiniBand VAPI interface based on the MPICH implementation

VAPI

Mellanox IB-Verbs API

Mellanox Infiniband Verbs Application Programming Interface

The `M' in MUG (part 2)



Message Passing Interface

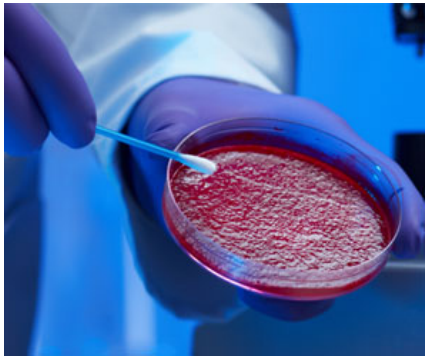
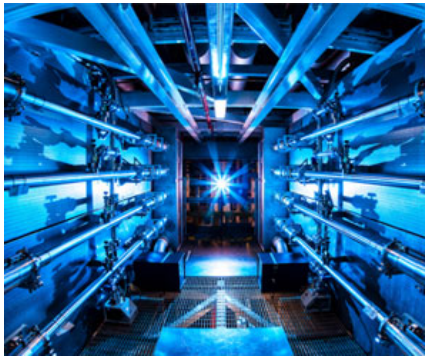
for

Mellanox Infiniband Verbs
Application Programming Interface

on

Chameleon

LLNL's mission is applying world-class science, technology, and engineering to national & global problems

Bio-Security	Counterterrorism	Defense	Energy
			
Intelligence	**Nonproliferation**	**Science**	**Weapons**
			

https://missions.llnl.gov

LLNL systems by purpose

Capability
Capacity
Visualization
Serial

System	Top500 Rank	Program	Manufacture / Model	OS	Inter-connect	Cores	Memory (GB)	Peak TFLOP/s
Unclassified Network (OCF)								
Vulcan	33	ASC+M&IC+HPCIC	IBM BGQ	RHEL/CNK	5D Torus	393,216	393,216	5,033.2
Lassen		ASC+M&IC	IBM P9	RHEL	2x IB EDR	30,096	218,880	19,886.0
Quartz	63	ASC+M&IC	Penguin	TOSS	Omni-Path	96,768	344,064	3251.4
Pascal		ASC+M&IC	Penguin	TOSS	IB EDR	5,868	41,728	1,700
RZTopaz		ASC	Penguin	TOSS	Omni-Path	27,648	98,304	929.0
RZManta		ASC	IBM P8	RHEL	IB EDR	720	11,520	597.6
Ray		ASC+M&IC	IBM P8	RHEL	IB EDR	1,080	17,280	896.4
RZAnsel		ASC	IBM P9	RHEL	2x IB EDR	2,376	17,280	1570.0
Catalyst		ASC+M&IC	Cray	TOSS	IB QDR	7,776	41,472	149.3
Syrah		ASC+M&IC	Cray	TOSS	IB QDR	5,184	20,736	107.8
Surface		ASC+M&IC	Cray	TOSS	IB FDR	2,592	41,500	451.9
Borax		ASC+M&IC	Penguin	TOSS	N/A	1,728	6,144	58.1
RZTrona		ASC	Penguin	TOSS	N/A	1,728	6,144	58.1
OCF Totals	**Systems**	**13**						**34,688.8**
Classified Network (SCF)								
Pinot(TLCC2, SNSI)		M&IC	Appro	TOSS	IB QDR	2,592	10,368	53.9
Sequoia	8	ASC	IBM BGQ	RHEL/CNK	5D Torus	1,572,864	1,572,864	20132.7
Sierra*	3	ASC	IBM P9	RHEL	2x IB EDR	190,080	1,382,400	125626.0
Zin (TLCC2)	437	ASC	Appro	TOSS	IB QDR	46,656	93,312	961.1
Jade+Jadeita	64	ASC	Penguin	TOSS	Omni-Path	96,768	344,064	3251.4
Mica		ASC	Penguin	TOSS	Omni-Path	13,824	49,152	464.5
Shark		ASC	IBM	RHEL	IB EDR	720	11,520	597.6
Max		ASC	Appro	TOSS	IB FDR	5,184	82,944	107.8
Agate		ASC	Penguin	TOSS	N/A	1,728	6,144	58.1
SCF Totals	Systems	9						**151,253.1**
Combined Totals		**22**						**185,941.9**

System	Top500 Rank	Program	Manufacture / Model	OS	Inter-connect	Cores	Memory (GB)	Peak TFLOP/s
Unclassified Network (OCF)								
Vulcan	33	ASC+M&IC+HPCIC	IBM BGQ	RHEL/CNK	5D Torus	393,216	393,216	5,033.2
Lassen		ASC+M&IC	IBM P9	RHEL	2x IB EDR	30,096	218,880	19,886.0
Quartz	63	ASC+M&IC	Penguin	TOSS	Omni-Path	96,768	344,064	3251.4
Pascal		ASC+M&IC	Penguin	TOSS	IB EDR	5,868	41,728	1,700
RZTopaz		ASC	Penguin	TOSS	Omni-Path	27,648	98,304	929.0
RZManta		ASC	IBM P8	RHEL	IB EDR	720	11,520	597.6
Ray		ASC+M&IC	IBM P8	RHEL	IB EDR	1,080	17,280	896.4
RZAnsel		ASC	IBM P9	RHEL	2x IB EDR	2,376	17,280	1570.0
Catalyst		ASC+M&IC	Cray	TOSS	IB QDR	7,776	41,472	149.3
Syrah		ASC+M&IC	Cray	TOSS	IB QDR	5,184	20,736	107.8
Surface		ASC+M&IC	Cray	TOSS	IB FDR	2,592	41,500	451.9
Borax		ASC+M&IC	Penguin	TOSS	N/A	1,728	6,144	58.1
RZTrona		ASC	Penguin	TOSS	N/A	1,728	6,144	58.1
OCF Totals	**Systems**	**13**						**34,688.8**
Classified Network (SCF)								
Pinot(TLCC2, SNSI)		M&IC	Appro	TOSS	IB QDR	2,592	10,368	53.9
Sequoia	8	ASC	IBM BGQ	RHEL/CNK	5D Torus	1,572,864	1,572,864	20132.7
Sierra*	3	ASC	IBM P9	RHEL	2x IB EDR	190,080	1,382,400	125626.0
Zin (TLCC2)	437	ASC	Appro	TOSS	IB QDR	46,656	93,312	961.1
Jade+Jadeita	64	ASC	Penguin	TOSS	Omni-Path	96,768	344,064	3251.4
Mica		ASC	Penguin	TOSS	Omni-Path	13,824	49,152	464.5
Shark		ASC	IBM	RHEL	IB EDR	720	11,520	597.6
Max		ASC	Appro	TOSS	IB FDR	5,184	82,944	107.8
Agate		ASC	Penguin	TOSS	N/A	1,728	6,144	58.1
SCF Totals	Systems	9						**151,253.1**
Combined Totals		**22**						**185,941.9**

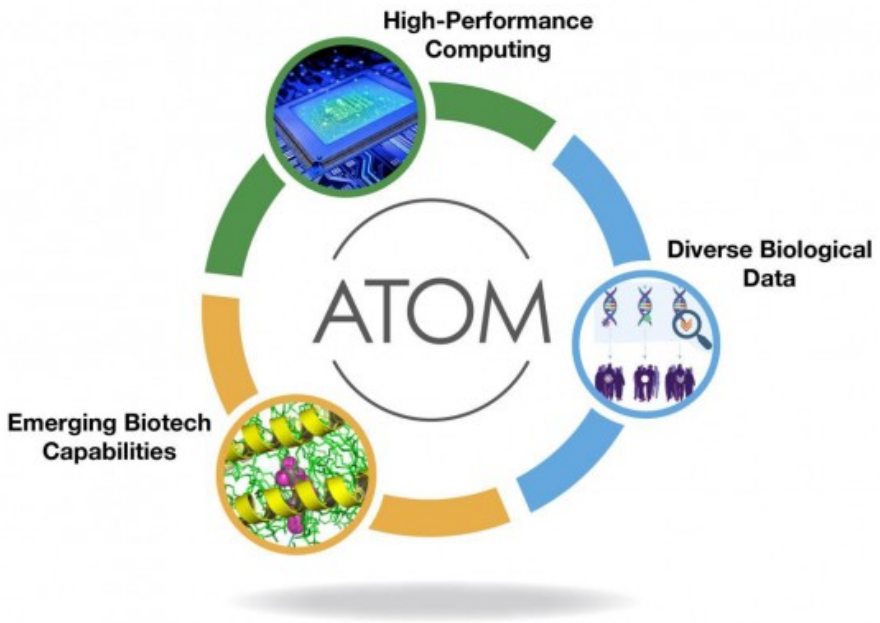
Science with MVAPICH

Public-private consortium aims to cut preclinical cancer drug discovery from six years to just one

Accelerating Therapeutics for Opportunities in Medicine (ATOM)

Lawrence Livermore National Lab, Frederick National Laboratory for Cancer Research, GSK and University of California, San Francisco

ATOM will develop, test and validate a multidisciplinary approach to drug discovery in which modern science, technology and engineering, supercomputing simulations, data science and artificial intelligence are highly integrated into a single drug-discovery platform that can ultimately be shared with the drug development community at large.



https://www.llnl.gov/news/public-private-consortium-aims-cut-preclinical-cancer-drug-discovery-six-years-just-one

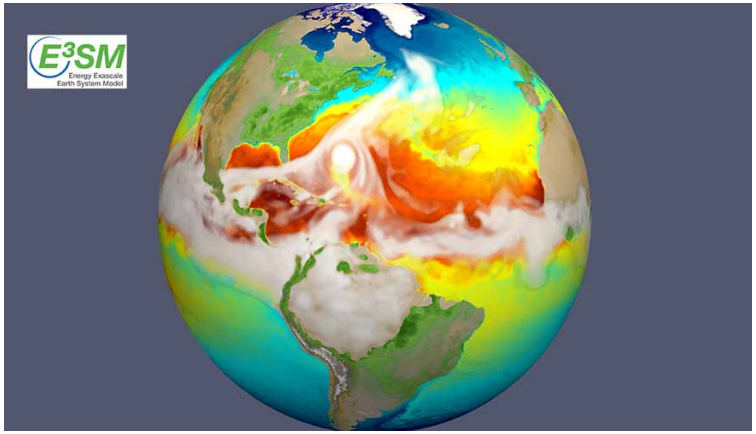
https://atom.cancer.gov/

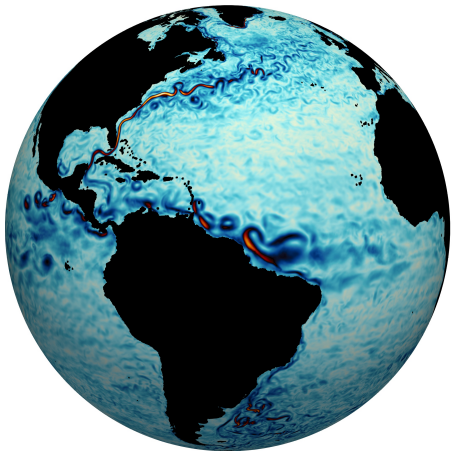
New exascale system for earth simulation

Energy Exascale Earth System Model (E3SM)

To address the diverse critical factors impacting the U.S. energy sector, the E3SM project is dedicated to answering three overarching scientific questions that drive its numerical experimentation initiatives:

- Water Cycle: How does the hydrological cycle interact with the rest of the human-earth system on local to global scales to determine water availability and water cycle extremes?
- Biogeochemistry: How do biogeochemical cycles interact with other earth system components to influence the energy sector?
- Cryosphere Systems: How do rapid changes in cryosphere (continental and ocean ice) systems evolve with the earth system, and contribute to sea-level rise and increased coastal vulnerability?"



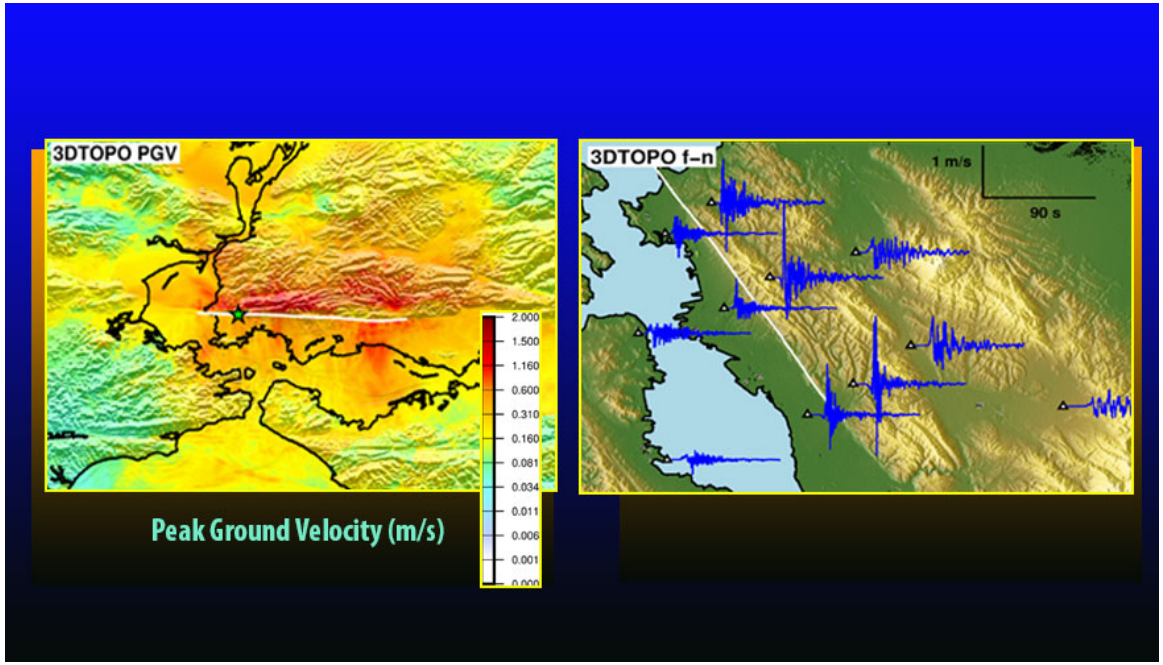


https://www.llnl.gov/news/new-exascale-system-earth-simulation
https://youtu.be/8Df96rx3i9g

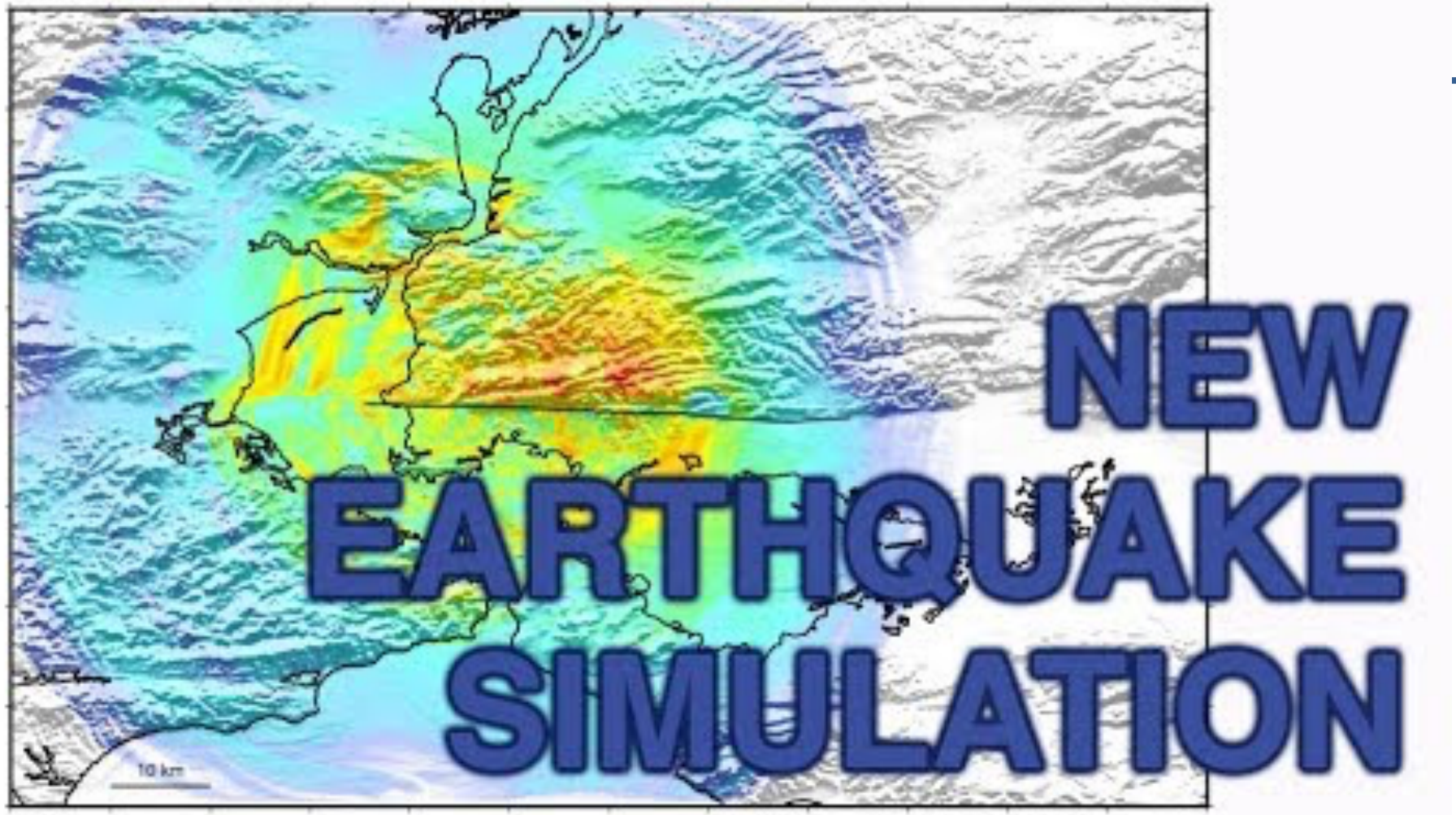
Hayward fault earthquake simulations increase fidelity of ground motions

In the next 30 years, there is a one-in-three chance that the Hayward fault will rupture with a 6.7 magnitude or higher earthquake, according to the United States Geologic Survey (USGS). Such an earthquake will cause widespread damage to structures, transportation and utilities, as well as economic and social disruption in the East Bay.

Lawrence Livermore (LLNL) and Lawrence Berkeley (LBNL) national laboratory scientists have used some of the world's most powerful supercomputers to model ground shaking for a magnitude (M) 7.0 earthquake on the Hayward fault and show more realistic motions than ever before. The research appears in *Geophysical Research Letters*.



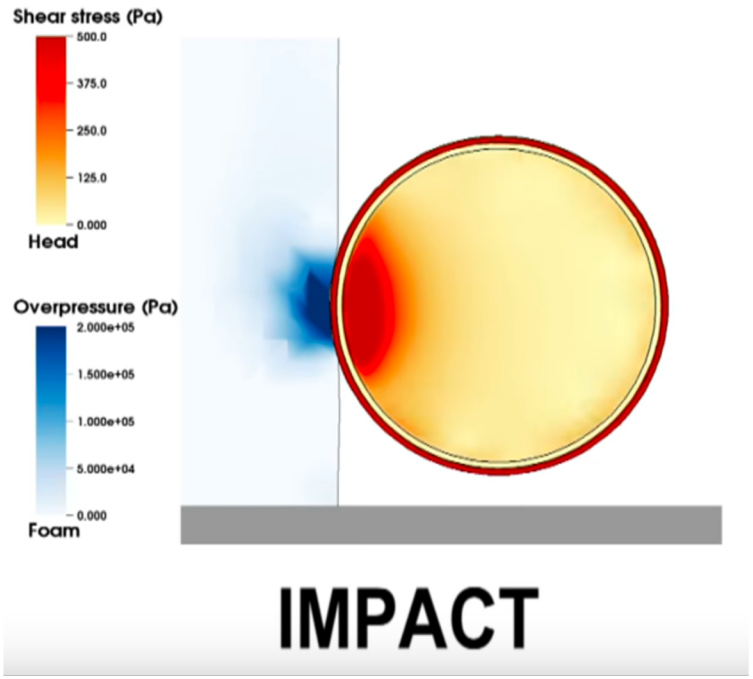
https://www.llnl.gov/news/hayward-fault-earthquake-simulations-increase-fidelity-ground-motions
https://www.llnl.gov/news/exascale-motion-earthquake-risks
https://youtu.be/nB1XOo_-uzU


10 km
NEW
EARTHQUAKE
SIMULATION

Concussion study may ‘change the game’

Researchers have identified evidence of early chronic traumatic encephalopathy (CTE) brain pathology after head impact -- even in the absence of signs of concussion. Early indicators of CTE pathology not only persisted long after injury but also spread through the brain, providing the best evidence to date that head impact, not concussion, causes CTE.

The findings, published online in the journal *Brain*, help to explain why approximately 20 percent of athletes with CTE never suffered a diagnosed concussion. The findings were based on analysis of human brains from teenagers with recent head injury, animal experiments that recreate sports-related head impact and military-related blast exposure, and computational models of the skull and brain during these injuries.



https://www.llnl.gov/news/concussion-study-may-‘change-game’
https://youtu.be/XpOlWR0Kf1k

Next round for UN climate change report begins

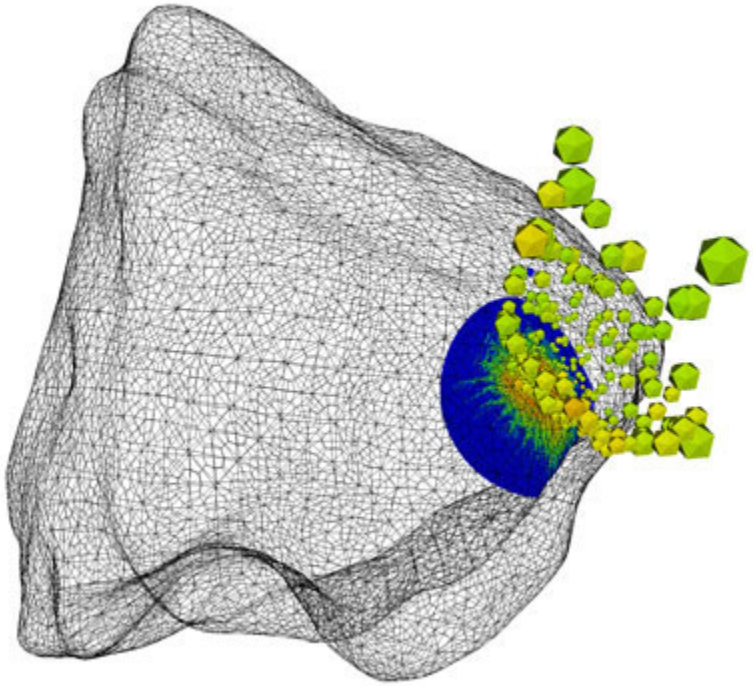
The seven-year cycle of scientific assessment driven by the United Nations Intergovernmental Panel on Climate Change (UN-IPCC) has begun.

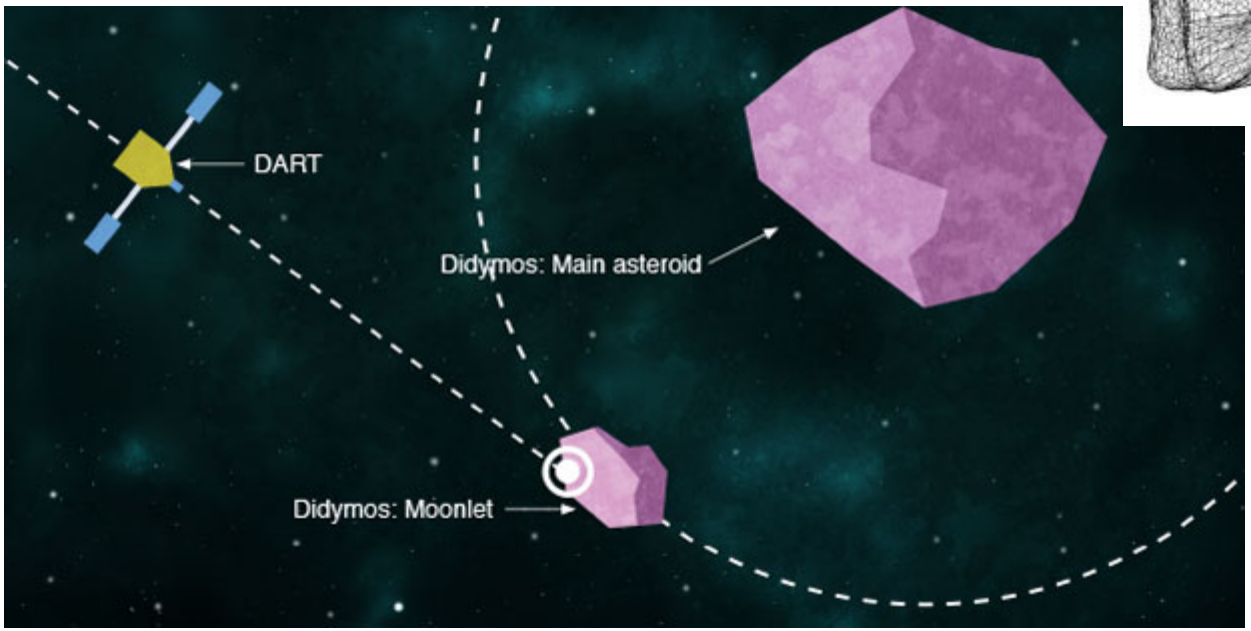
This contribution continues a four-decade legacy of Lawrence Livermore National Laboratory (LLNL) involvement in the IPCC, with scientific leadership provided through all five of the IPCC assessments dating back to the initial report, published in 1990.

https://www.llnl.gov/news/next-round-un-climate-change-report-begins

Asteroid Deflection

- Model how asteroid trajectory changes after impact from spacecraft
- DART: Joint mission with JHU APL to launch a test in 2020





http://youtu.be/xXCxMeZ-yQo

https://str.llnl.gov/december-2016/syal

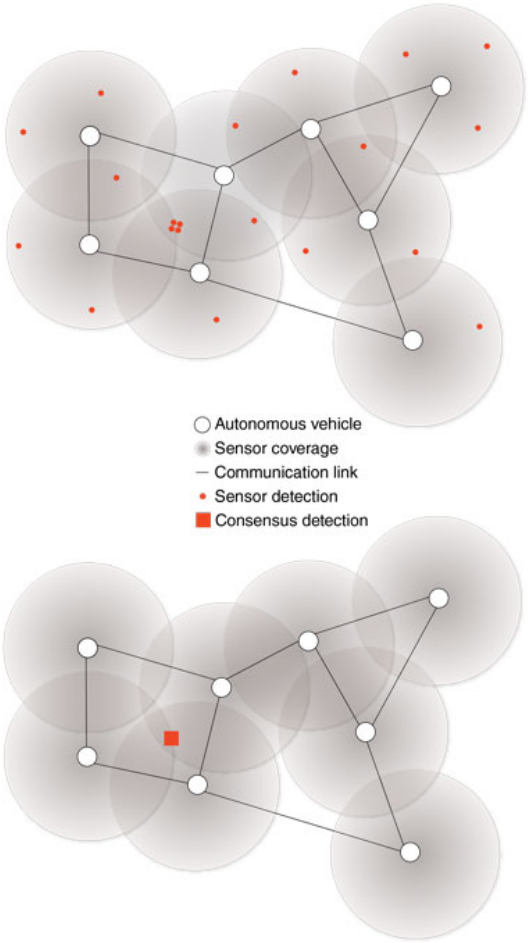
UCRL-TR-52000-16-12

MVAPICH
enables science that
changes the world

Collaborative Autonomy

“Building a Network of Collaborative Autonomous Machines”

- Swarm of distributed drones with sensors cooperate to arrive at decision
- Example: locate people in burning building
- MPI used to model convergence time on comm graphs with varying network properties (packet loss, latency, bandwidth)

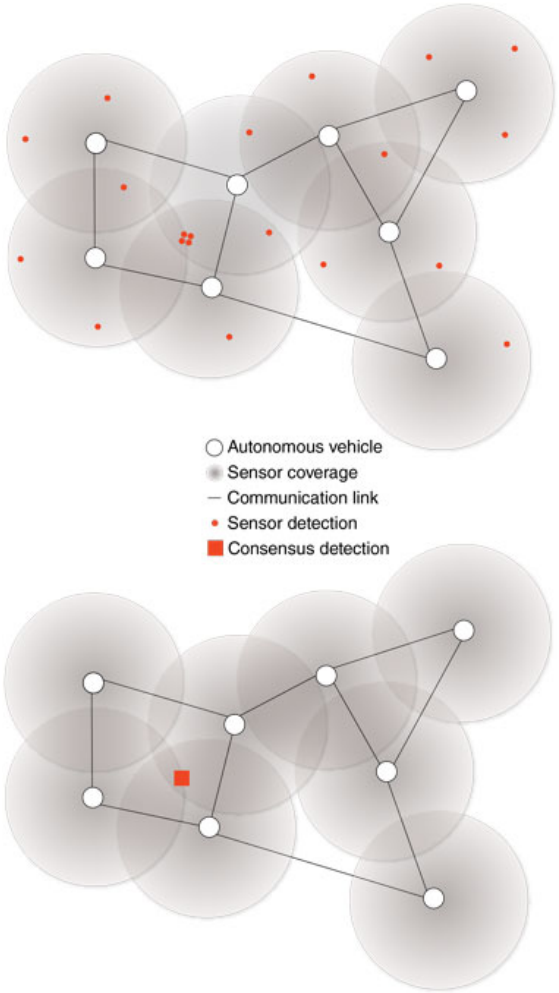


https://str.llnl.gov/2018-06/beer

UCRL-TR-52000-18-6, Distribution Category UC-99, June 2018

ns-3: discrete-event network simulator for Internet systems

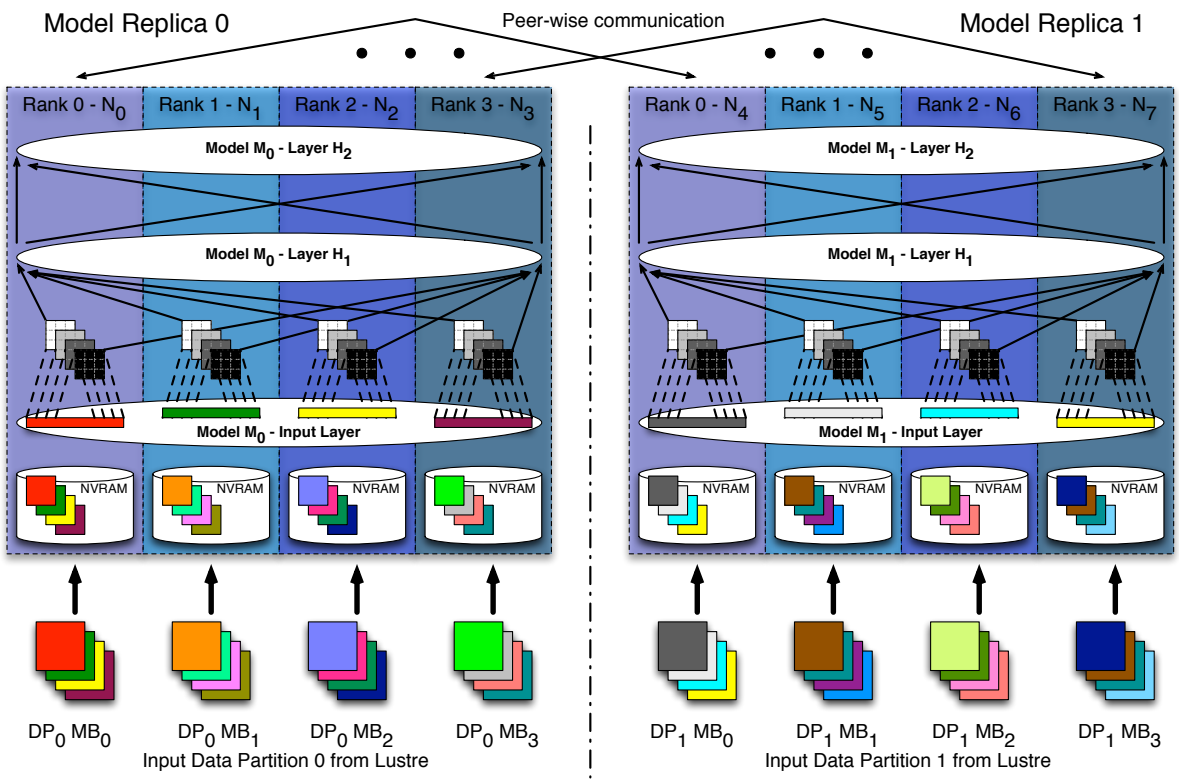
- Implementation
  - https://www.nsnam.org
  - Parallel version heavy pt-2-pt and Allreduce
  - Point-to-point is irregular, non-blocking
- Largest LLNL jobs
  - Millions of simulated nodes in network
  - Scaled to hundreds of compute nodes and thousands of processes
- Requests
  - Efficient Allreduce using custom datatype/operations (SUM with MIN)
  - Non-blocking Allreduce with true overlap
  - Looking to over decompose to improve comp/comm overlap
  - Misses the C++ MPI bindings



HPC for Deep Learning

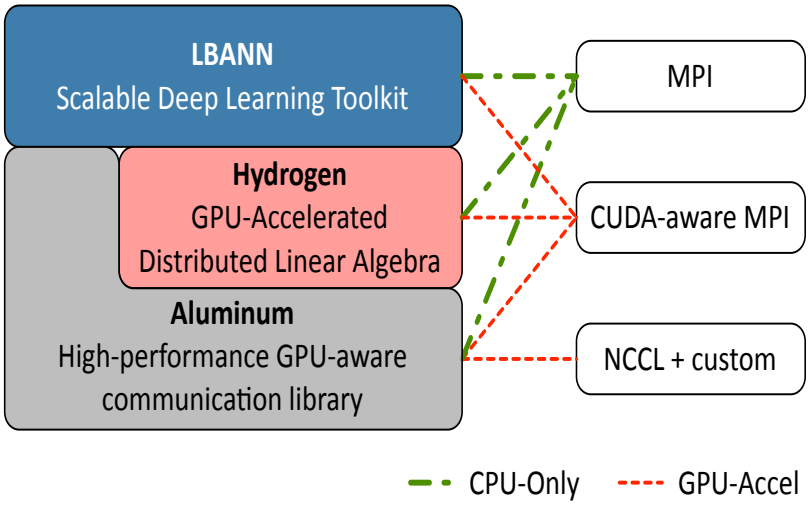
LBANN: Livermore Big Artificial Neural Network Toolkit

- Deep Neural Network training / classification
  - Optimized distributed memory algorithm
  - Optimized asynchronous all-reduce library
  - Train large networks fast
  - Optimize for strong & weak scaling
- Unique HPC resources at scale
  - InfiniBand or Omnipath interconnect
    - low latency / high cross section bandwidth
  - Tightly-coupled GPU accelerators
  - Node-local NVRAM
  - High bandwidth Parallel File System
  - State-of-the art distributed linear algebra library
- Open source under Apache license
  - https://github.com/LLNL/lbann



LBANN is pushing supported limits of MPI, threads, and GPUs

- Improve Allreduce algorithms for user-defined datatypes/ops
  - Allreduce on compressed data
- High-precision accumulate for low-precision inputs
  - e.g., 16-bit internal Allreduce on 8-bit input / output data
- True non-blocking Allreduce and pt-2-pt
  - Overlap messages with backprop steps
- NCCL-like performance from MPI collectives
- Improve support for large-bandwidth messages
- Combining threaded MPI with GPUs non-trivial
- Need to associate GPU context with MPI comm
  - Some threads make MPI calls, but don't have GPU context
  - CUDA callbacks may invoke MPI, which kicks progress engine that makes CUDA call
- Need better methods to send/wait for messages from GPU kernels
  - What is state of GPUDirect Async?



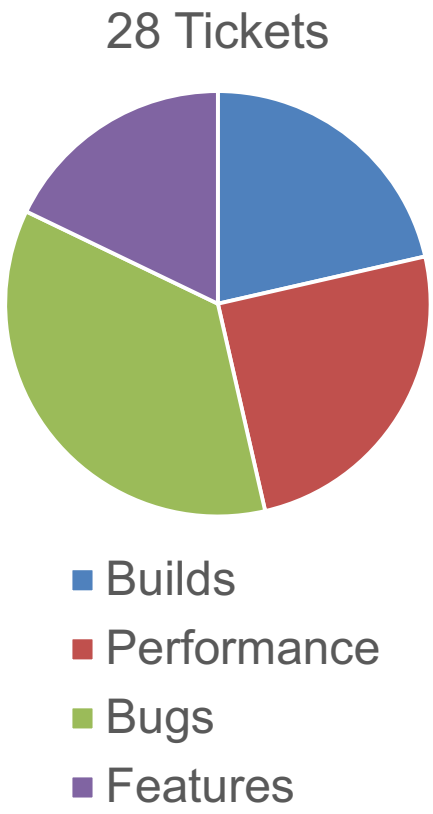
C++ / MPI + OpenMP / CUDA

Open-sourced on github.com

Improving MVAPICH to Enable Science

Tweaking MVAPICH for LLNL users at rate of one ticket per week via X-ScaleSolutions

- 28 tickets created over 7 months
- Tune MV2 for LLNL hardware configurations
  - Networks: Mellanox, Intel Omnipath, Shared Memory
  - CPUs: Intel, IBM Power, ARM
  - GPUs: NVIDIA (with GPUDirect RDMA)
- Create RPMs for LLNL systems for all supported CPUs, networks, and compilers
  - MV2-X
  - MV2-GDR
- Fix functionality and key performance bugs
- Small feature adds and configuration changes



mpiBench added to MVAPICH testing to detect performance problems

- mpiBench measures collective performance
- Processing script associates cost of (collective, node, ppn, msg size) tuple
- Self consistency checks
  - Verify that block vs cyclic is close
  - Avoid performance regression across updates, e.g., MV2-2.2 to MV2-2.3
- Flag violations of “rules” across collectives, e.g.
  - for a given (node count, ppn, msg size) Allreduce < Reduce + Bcast

```
my @rules = (
  "Barrier < Bcast",
  "Bcast < Gather",
  "Gather < Allgather",
  "Allgather < Alltoall",

  "Barrier < Allreduce",
  "Reduce < Allreduce",
  "Allreduce < Gather",

  "Gather = Scatter",
  "Scan = Allreduce",
  "Scan = Exscan",

  "Allreduce < Reduce + Bcast",
  "Allgather < Gather + Bcast",
);
```

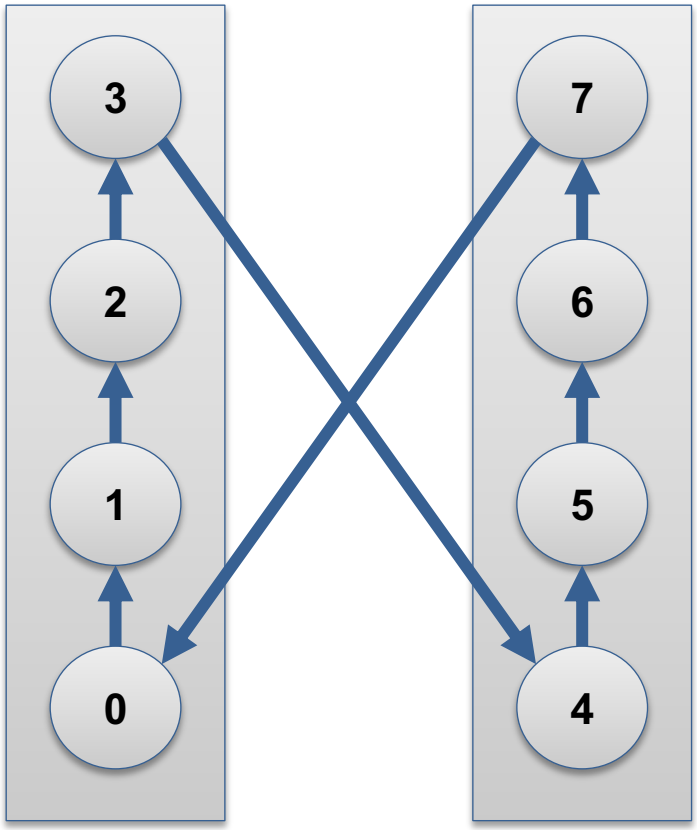
MVAPICH team improved Alltoall for important pf3d use case in MV2-2.3

- pf3d executes frequent on-node FFTs
- User reported that MV2-2.2 was slower than Intel MPI
- MPI_Alltoall identified as bottleneck
- MVAPICH team significantly improved on-node Alltoall

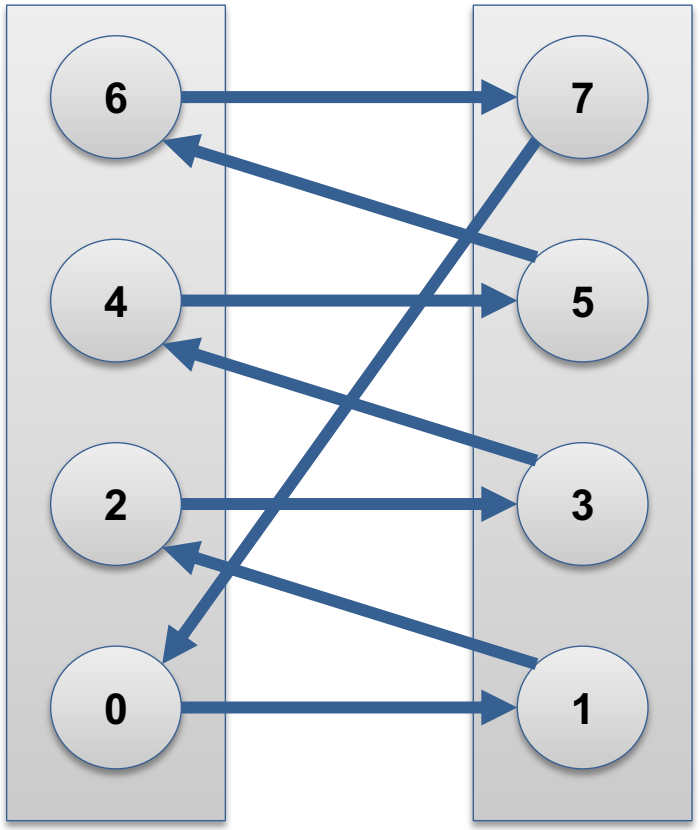
MPI_Alltoall speedup from 2.2 to 2.3

Alltoall	ppn	1	2	4	8	16	32
msgsize	nodes	1	1	1	1	1	1
8		0.353	0.836	0.732	1.076	1.899	1.797
16		0.420	0.481	0.873	1.354	1.690	1.879
32		0.351	0.846	1.056	1.402	2.139	3.396
64		0.378	0.716	0.881	1.179	1.606	2.386
128		0.195	0.683	0.874	1.089	1.419	2.123
256		0.164	0.683	0.818	1.000	1.303	1.542
512		0.160	0.695	0.794	1.001	1.112	1.281
1K		0.150	0.728	0.857	0.821	0.769	0.823
2K		0.166	0.850	0.834	0.842	0.833	0.933
4K		0.204	0.884	1.000	1.023	1.035	1.100
8K		0.226	0.943	0.958	0.911	0.995	0.968
16K		0.375	0.566	1.045	1.155	0.993	0.554
32K		0.613	0.527	0.811	0.802	0.619	0.310
64K		0.652	0.786	1.328	1.225	1.379	1.430
128K		0.891	0.646	1.130	1.127	1.480	1.346
256K		0.813	0.830	1.382	0.997	1.694	1.398
512K		0.956	0.840	1.598	1.244	1.883	1.410
1M		0.864	0.863	1.749	1.814	2.011	1.425

Ring algorithm for large messages can send data over network multiple times if based on MPI rank



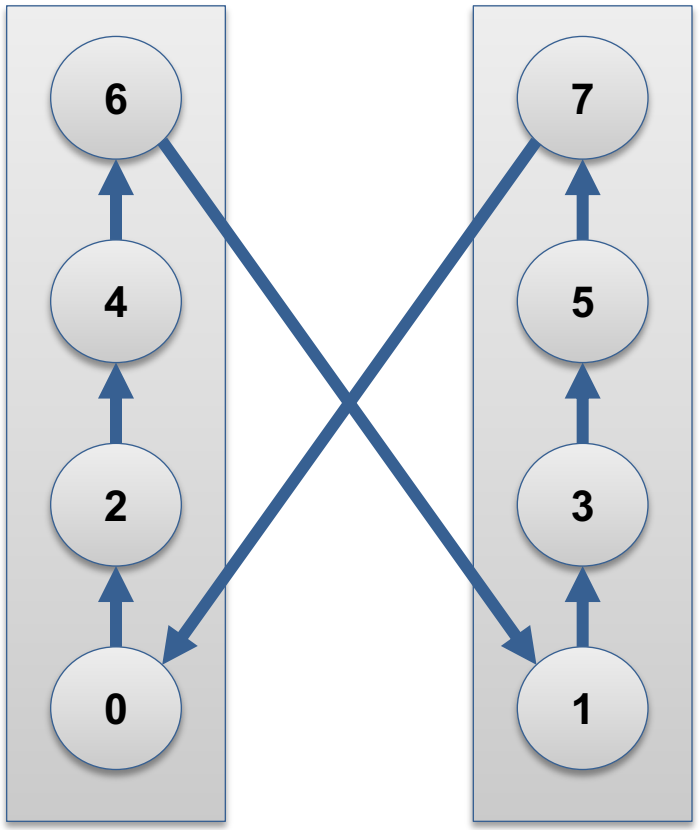
Block



Cyclic

Large message Allreduce, Allgather, Reduce_scatter up to 3 times faster

- Store ordered list of ranks across nodes in communicator structure
- Optimizes ring-based algorithms
  - Allgather
  - Reduce_scatter
  - Allreduce
- Delivers similar performance regardless of rank-to-node mapping
- Improves large message performance up to 3x
- Added in MVAPICH2-2.3
- Thanks Danielle Sikich



How to pronounce “MVAPICH”?

Answer #2 on MVAPICH Frequently Asked Questions

MVAPICH is pronounced as ``em-vah-pich".

A solution to clear up confusion among different MVAPICH flavors

Original Name	New Name	Pronounced as
MVAPICH2	MVAPICH	em – va – pitch
MVAPICH2-GDR	MVAPICH	em – vah – pitch
MVAPICH2-X	MVAPICH	M – V – A – P – I – C – H
MVAPICH2-Virt	MVAPICH	em – vah – pitch’
MVAPICH2-EA	MVAPICH	em – va – pitch’

LLNL HPC (+ MVAPICH) improves our world

"What you're doing in a lot of different areas has the potential to change the world," Perry said. "The computational capacity, what you have the potential to do, is nothing less than world-changing. This Lab is going to be part of a story, it may not be 10 years from now—it may be sooner than that, of how people's lives really get affected in a positive way."

— Rick Perry, DOE Secretary
March 26, 2018

https://computation.llnl.gov/newsroom/doe-secretary-visits-llnl-livermore-computing-center

Thank you, MVAPICH!

Thank you, Message Passing Interface for Mellanox Infiniband Verbs Application Programming Interface for Chameleon (aka MVAPICH)!

Thanks to LLNL contributors to this talk

- Jonathan Allen
- Peter Barnes
- Luc Peterson
- Brian Van Essen
- Nikoli Dryden
- Tim Fahey
- Meg Epperly
- Kathryn Mohror
- Scott Futral

Lawrence Livermore
National Laboratory