Faster, Bigger, Better Science with MVAPICH

MVAPICH User's Group Meeting

August 20, 2015

Presented by Adam Moody

Lawrence Livermore National Laboratory

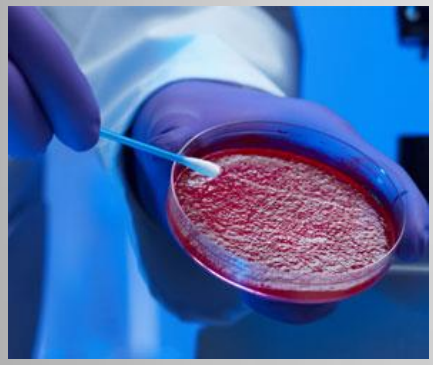
LLNL-PRES-676271

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

Who is LLNL and what do we do?

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	Nodes	Cores	Memory (GB)	Peak TFLOP/s
Unclassified Network (OCF)									
Vulcan	9	ASC+M&IC+HPCIC	IBM BGQ	RHEL/CNK	5D Torus	24,576	393,216	393,216	5,033.2
Sierra	417	M&IC	Dell	TOSS	IB QDR	1,944	23,328	46,656	243.7
Cab (TLCC2)	145	ASC+M&IC+HPCIC	Appro	TOSS	IB QDR	1,296	20,736	41,472	426.0
Ansel		M&IC	Dell	TOSS	IB QDR	324	3,888	7,776	43.5
RZMerl (TLCC2)		ASC+ICF	Appro	TOSS	IB QDR	162	2,592	5,184	53.9
RZZeus		M&IC	Appro	TOSS	IB DDR	267	2,144	6,408	20.6
Catalyst		ASC+M&IC	Cray	TOSS	IB QDR	324	7,776	41,472	149.3
Syrah		ASC+M&IC	Cray	TOSS	IB QDR	324	5,184	20,736	107.8
Surface		ASC+M&IC	Cray	TOSS	IB FDR	162	2,592	41,500	451.9
Aztec		M&IC	Dell	TOSS	N/A	96	1,152	4,608	12.9
Herd		M&IC	Appro	TOSS	IB DDR	9	256	1,088	1.6
OCF Totals	Systems	11							6,544.4
Classified Network (SCF)									
Pinot(TLCC2, SNSI)		M&IC	Appro	TOSS	IB QDR	162	2,592	10,368	53.9
Sequoia	3	ASC	IBM BGQ	RHEL/CNK	5D Torus	98,304	1,572,864	1,572,864	20132.7
Zin (TLCC2)	66	ASC	Appro	TOSS	IB QDR	2,916	46,656	93,312	961.1
Juno (TLCC)		ASC	Appro	TOSS	IB DDR	1,152	18,432	36,864	162.2
Muir		ICF	Dell	TOSS	IB QDR	1,296	15,552	31,104	168.0
Graph		ASC	Appro	TOSS	IB DDR	576	13,824	72,960	107.5
Max		ASC	Appro	TOSS	IB FDR	324	5,184	82,944	107.8
Inca		ASC	Dell	TOSS	N/A	100	1,216	5,120	13.5
SCF Totals	Systems	8							21,706.7
Combined Totals		19							28,251.1

System	Top500 Rank	Program	Manufacture/ Model	OS	Inter-connect	Nodes	Cores	Memory (GB)	Peak TFLOP/s
Unclassified Network (OCF)									
Vulcan	9	ASC+M&IC+HPCIC	IBM BGQ	RHEL/CNK	5D Torus	24,576	393,216	393,216	5,033.2
Sierra	417	M&IC	Dell	TOSS	IB QDR	1,944	23,328	46,656	243.7
Cab (TLCC2)	145	ASC+M&IC+HPCIC	Appro	TOSS	IB QDR	1,296	20,736	41,472	426.0
Ansel		M&IC	Dell	TOSS	IB QDR	324	3,888	7,776	43.5
RZMerl (TLCC2)		ASC+ICF	Appro	TOSS	IB QDR	162	2,592	5,184	53.9
RZZeus		M&IC	Appro	TOSS	IB DDR	267	2,144	6,408	20.6
Catalyst		ASC+M&IC	Cray	TOSS	IB QDR	324	7,776	41,472	149.3
Syrah		ASC+M&IC	Cray	TOSS	IB QDR	324	5,184	20,736	107.8
Surface		ASC+M&IC	Cray	TOSS	IB FDR	162	2,592	41,500	451.9
Aztec		M&IC	Dell	TOSS	N/A	96	1,152	4,608	12.9
Herd		M&IC	Appro	TOSS	IB DDR	9	256	1,088	1.6
OCF Totals	Systems	11							6,544.4
Classified Network (SCF)									
Pinot(TLCC2, SNSI)		M&IC	Appro	TOSS	IB QDR	162	2,592	10,368	53.9
Sequoia	3	ASC	IBM BGQ	RHEL/CNK	5D Torus	98,304	1,572,864	1,572,864	20132.7
Zin (TLCC2)	66	ASC	Appro	TOSS	IB QDR	2,916	46,656	93,312	961.1
Juno (TLCC)		ASC	Appro	TOSS	IB DDR	1,152	18,432	36,864	162.2
Muir		ICF	Dell	TOSS	IB QDR	1,296	15,552	31,104	168.0
Graph		ASC	Appro	TOSS	IB DDR	576	13,824	72,960	107.5
Max		ASC	Appro	TOSS	IB FDR	324	5,184	82,944	107.8
Inca		ASC	Dell	TOSS	N/A	100	1,216	5,120	13.5
SCF Totals	Systems	8							21,706.7
Combined Totals		19							28,251.1

Why MVAPICH?

Why MVAPICH?

- First MPI available for IB
- Reliable and proven
- Fastest for many users
- Familiarity with MPICH code base
- Acceptance of feedback and patches
- Good ties and communication with OSU

Livermore Computing / OSU: Successful history of collaboration

- Interns
  - Matt Koop
  - Hari Subramoni
  - Krishna Kandalla
  - Raghunath Rajachandrasekar
  - Sourav Chakraborty
- Compute resources
  - LC Collaborative Zone
  - Hyperion





Science with MVAPICH
Ex. 1

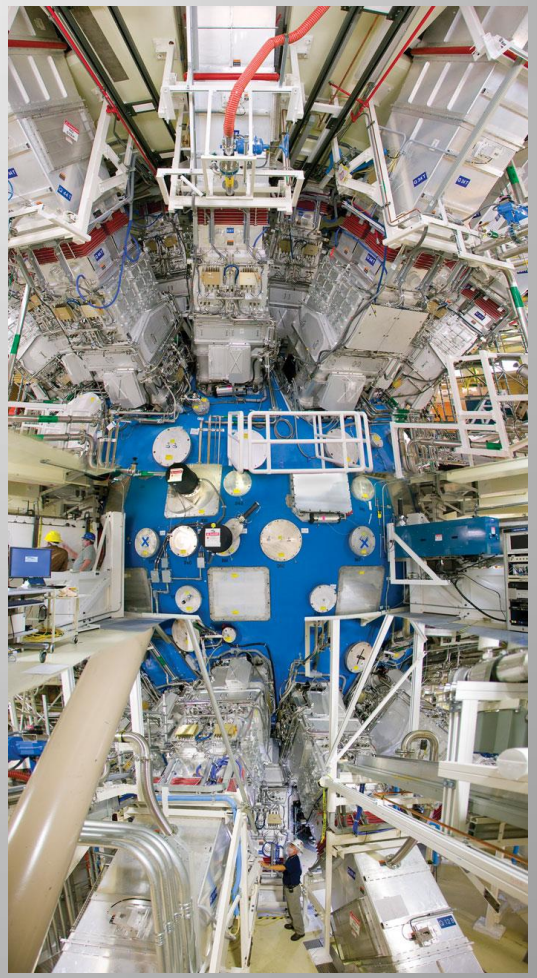
The National Ignition Facility: World's Highest Energy Laser

Bigger than 3 football fields

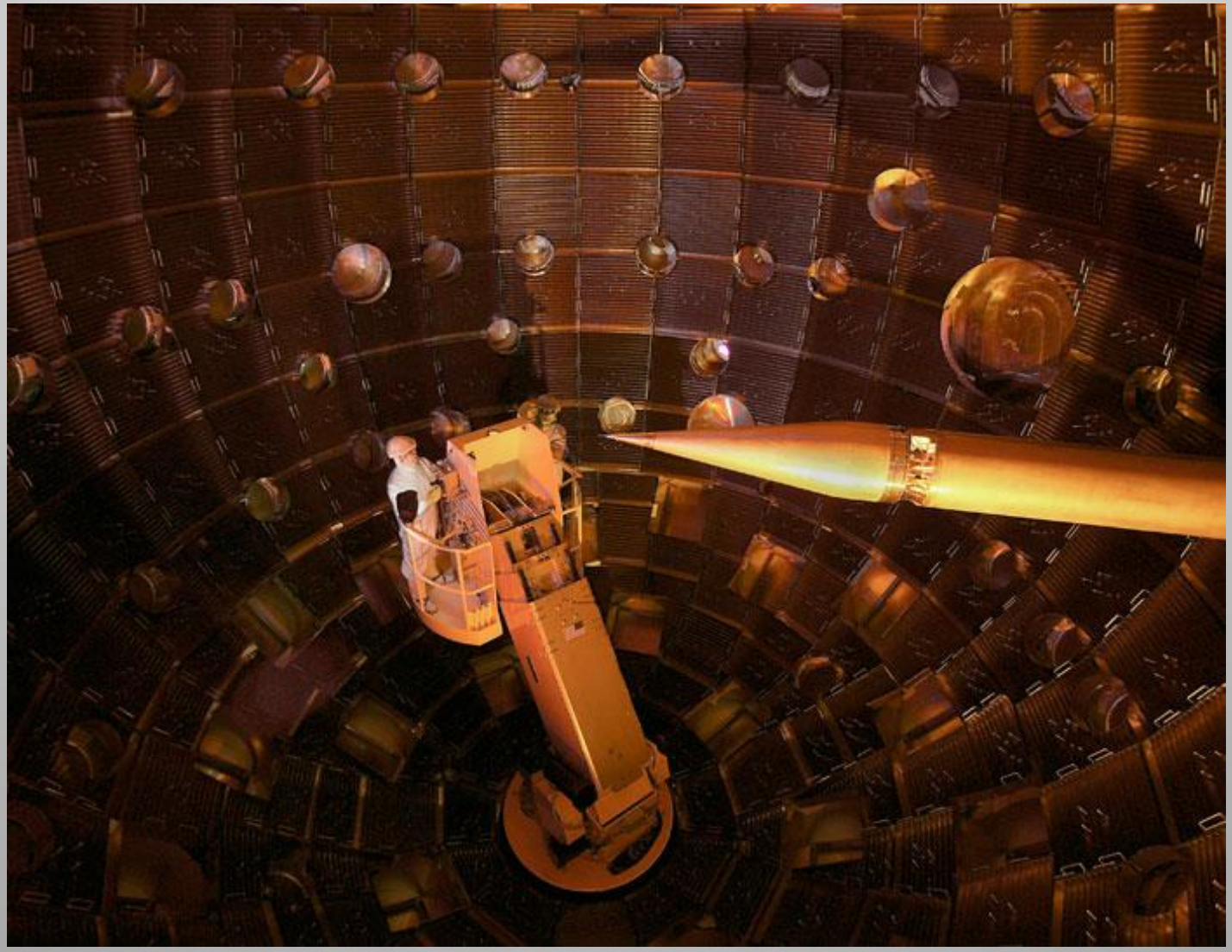
Laser is split and amplified

Most of the facility is the laser bay

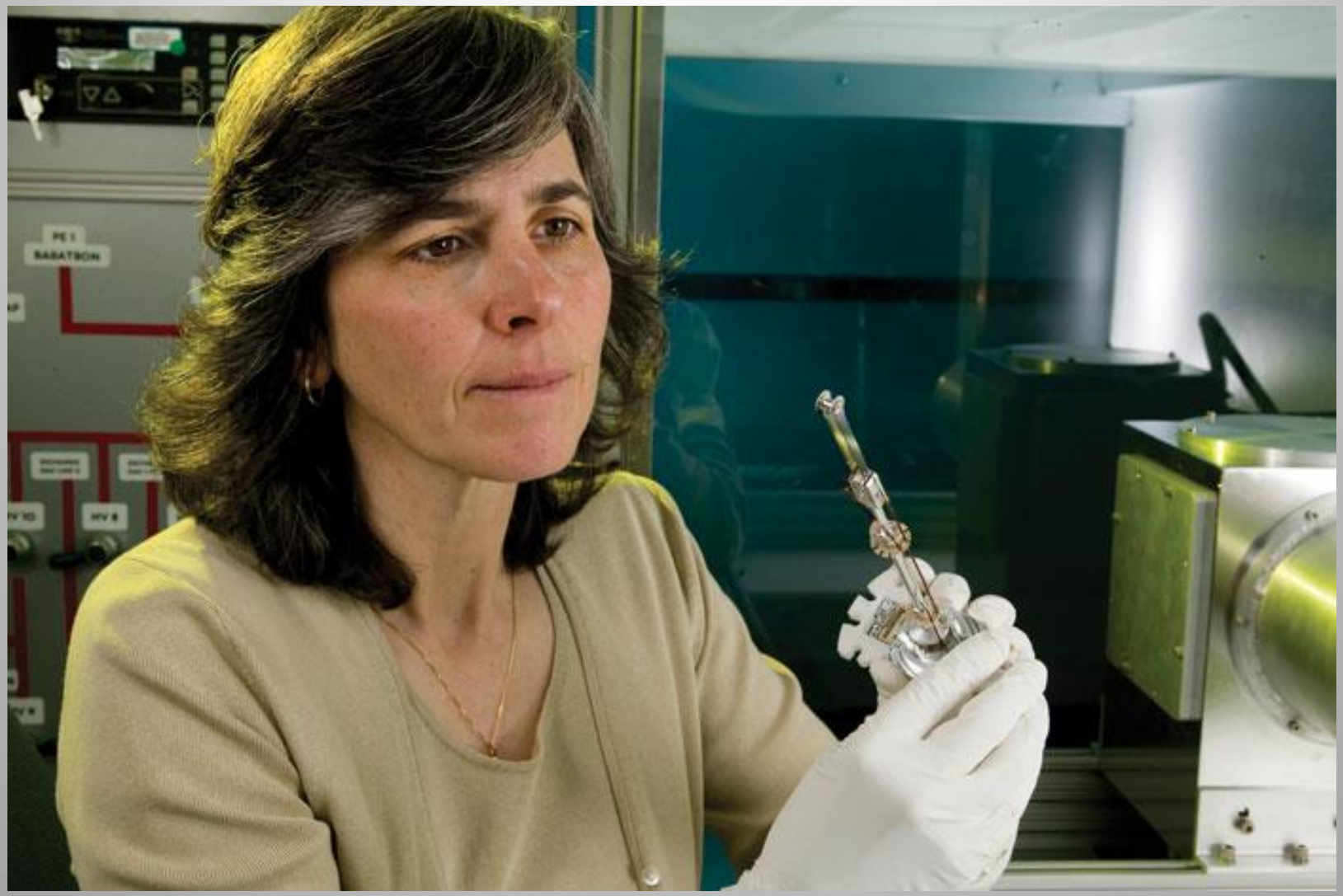
192 lasers directed to target chamber



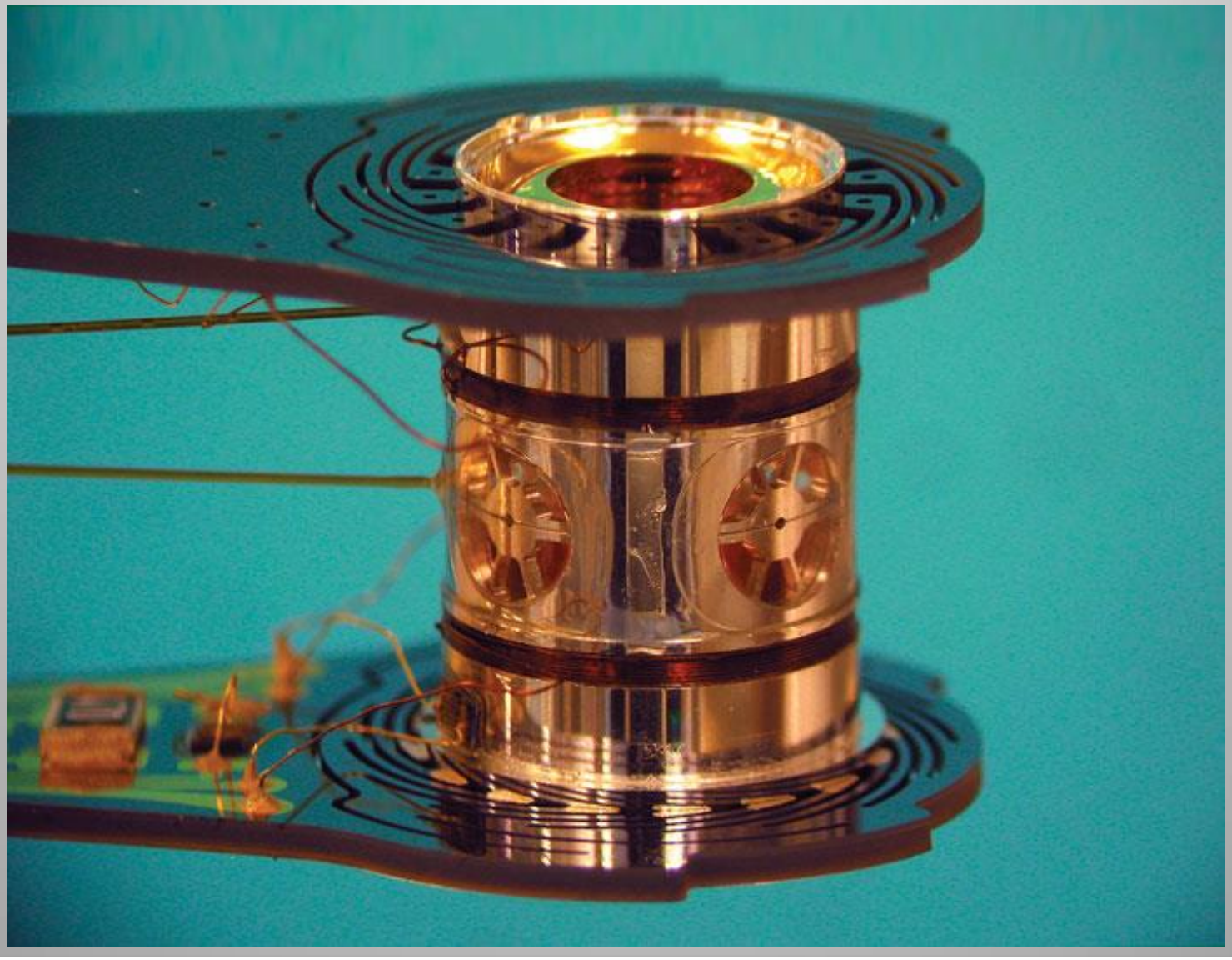
Target chamber is 30 feet in diameter



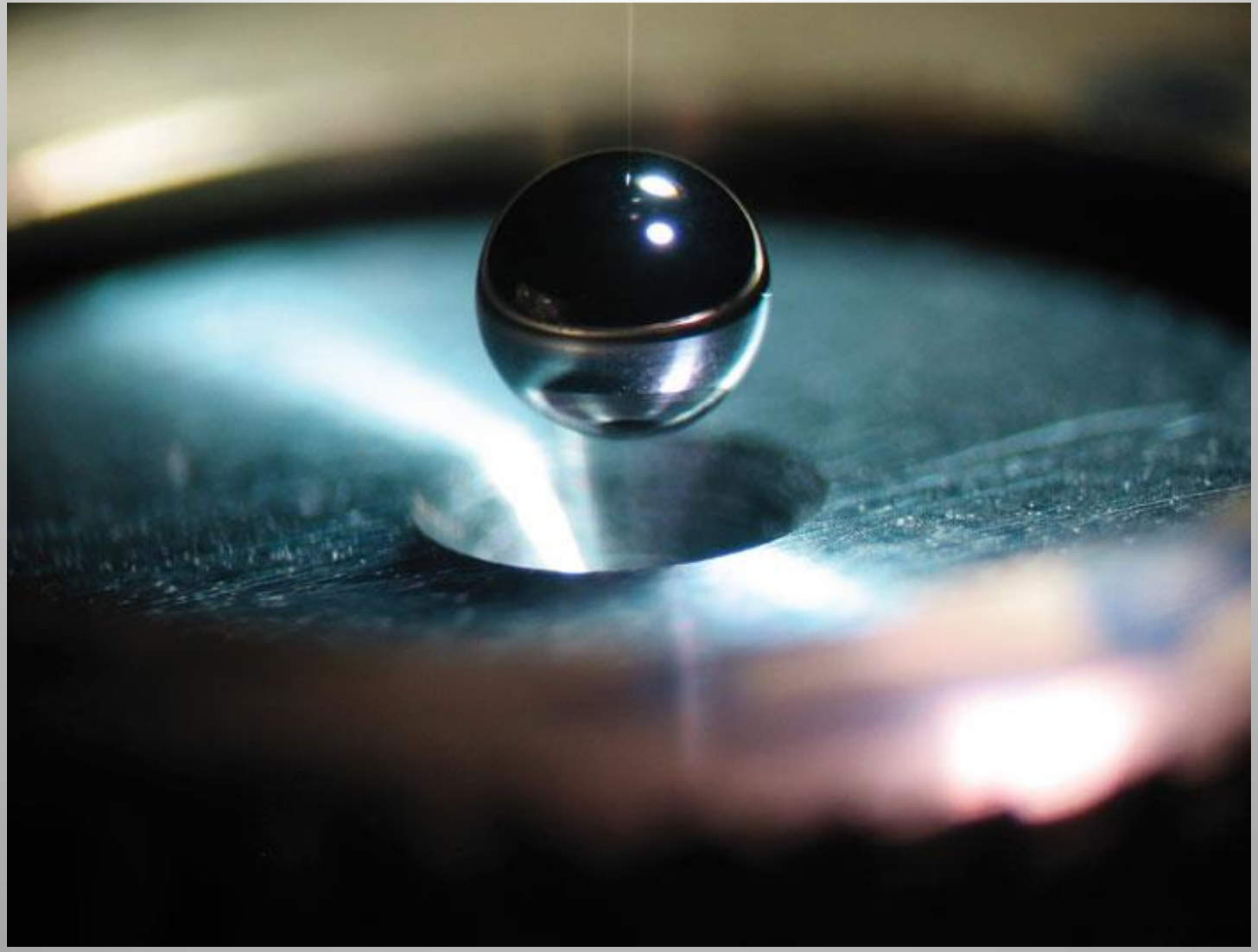
Target is tiny



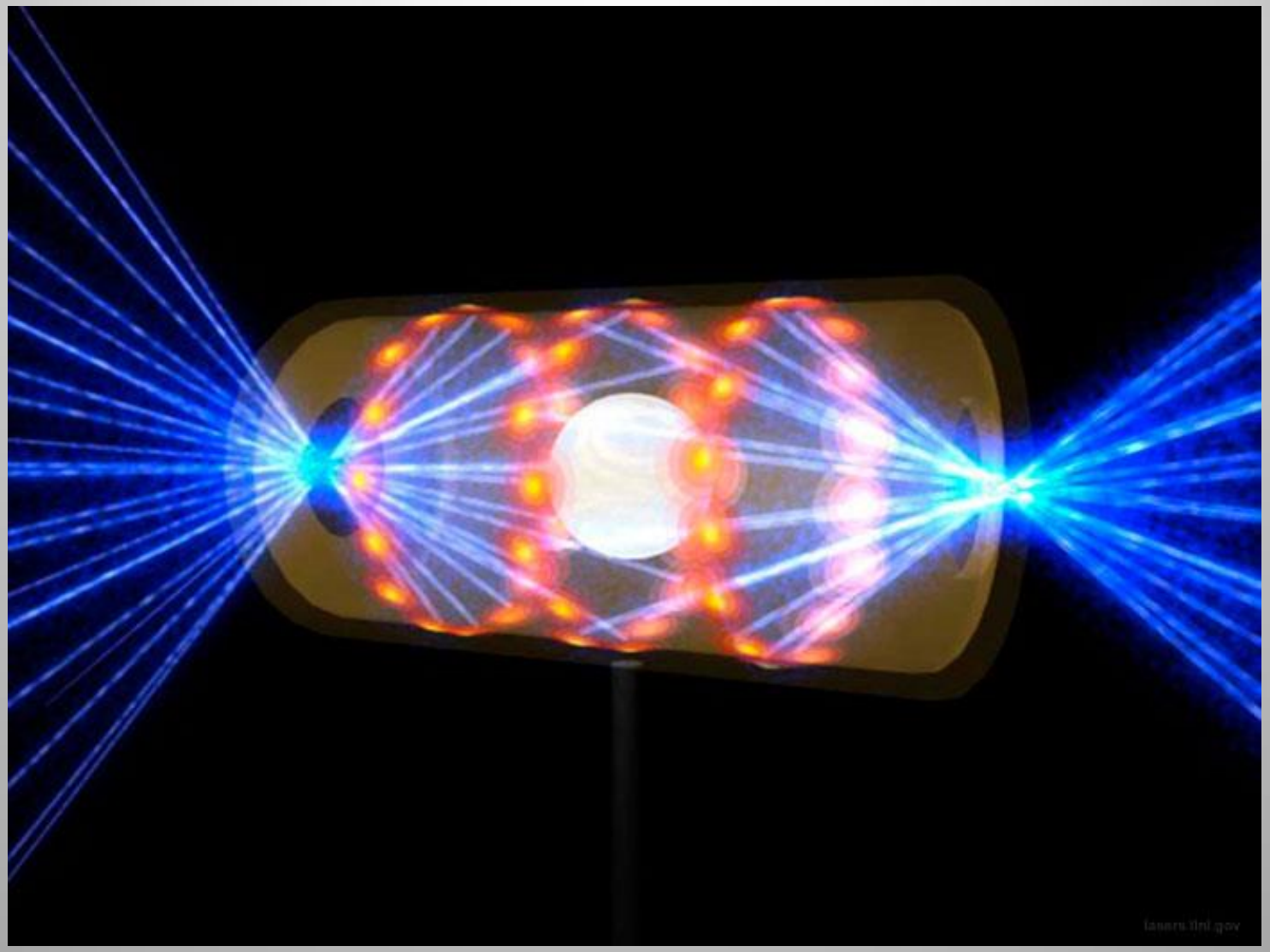
Gold cylinder called hohlraum



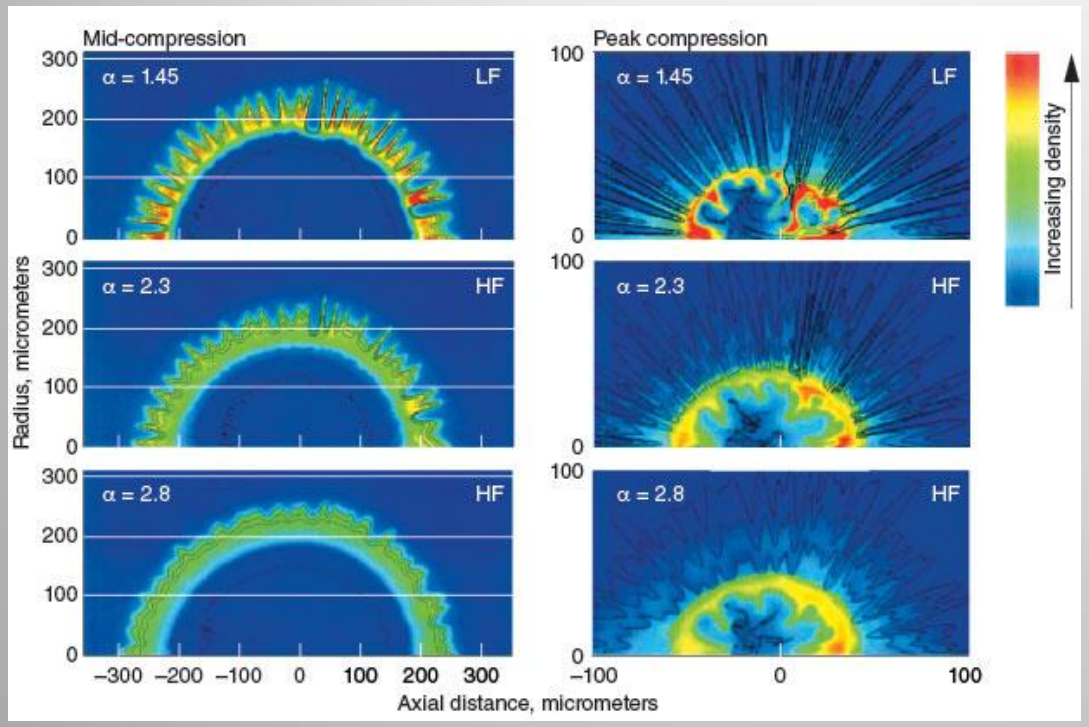
BB-sized ball of frozen hydrogen



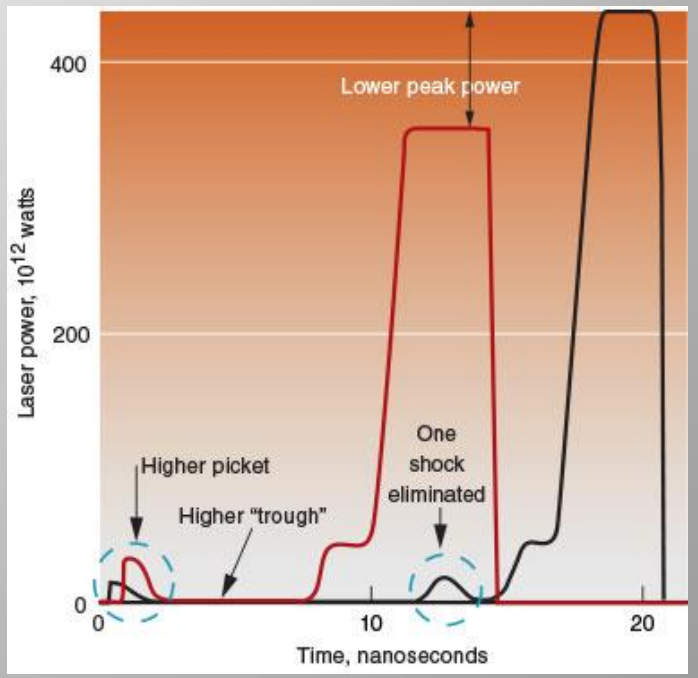
Lasers, x-rays, and implosion



Turbulent mixing of ablator shell limits fuel compression

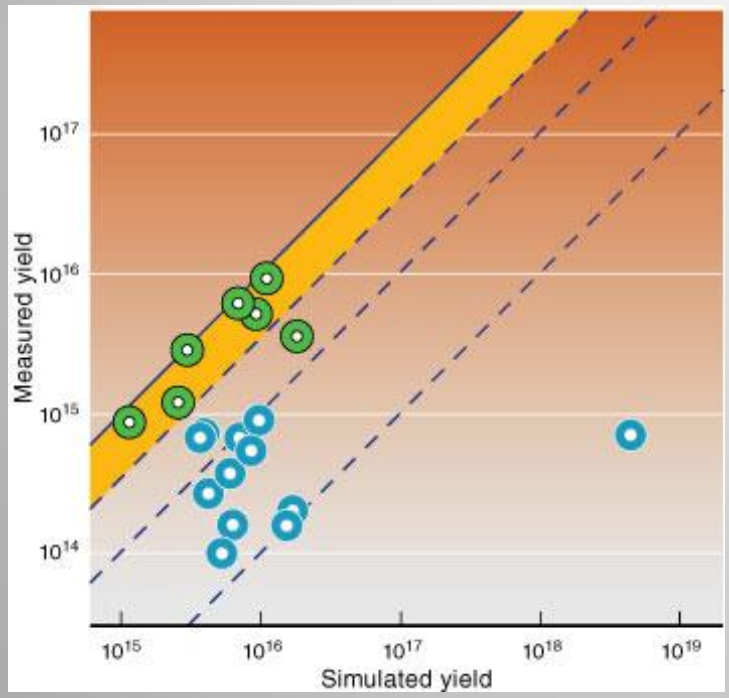


Consider reshaping laser pulse to reduce mixing



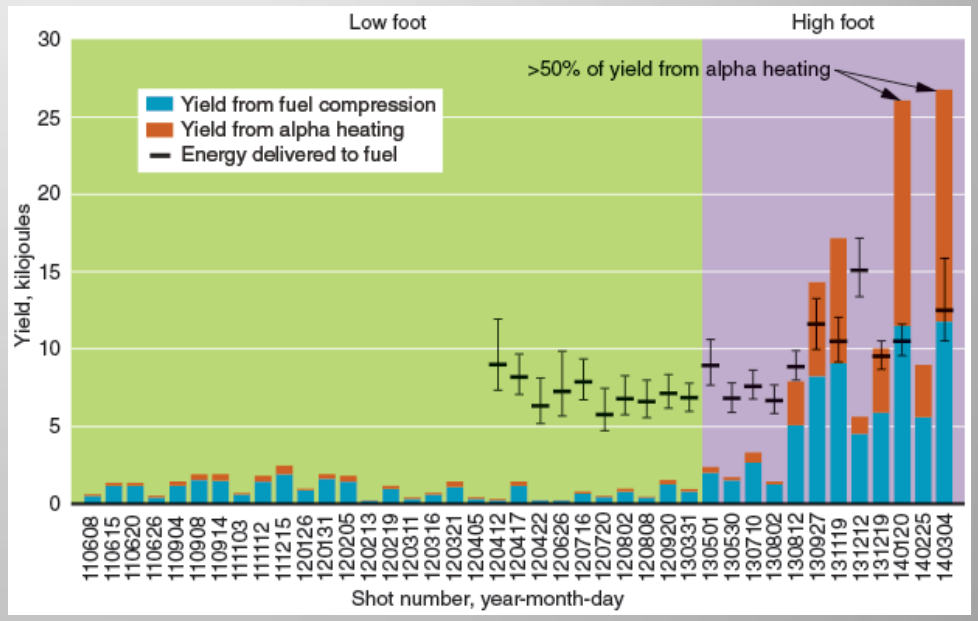
Computer simulations using MVAPICH support the theory

Shots with new pulse produce more energy than input from laser



Experiments agree exceptionally well with simulations

Major step forward on path to ignition



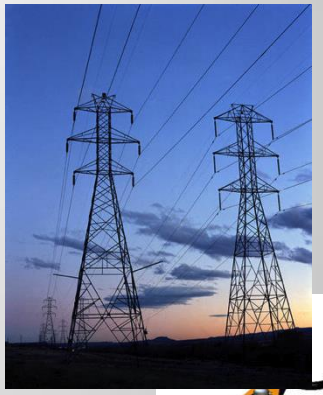
"Fuel gain exceeding unity in an inertially confined fusion implosion", Hurricane et. al., Nature, Feb 2014

https://str.llnl.gov/june-2014/hurricane

Science with MVAPICH
Ex. 2

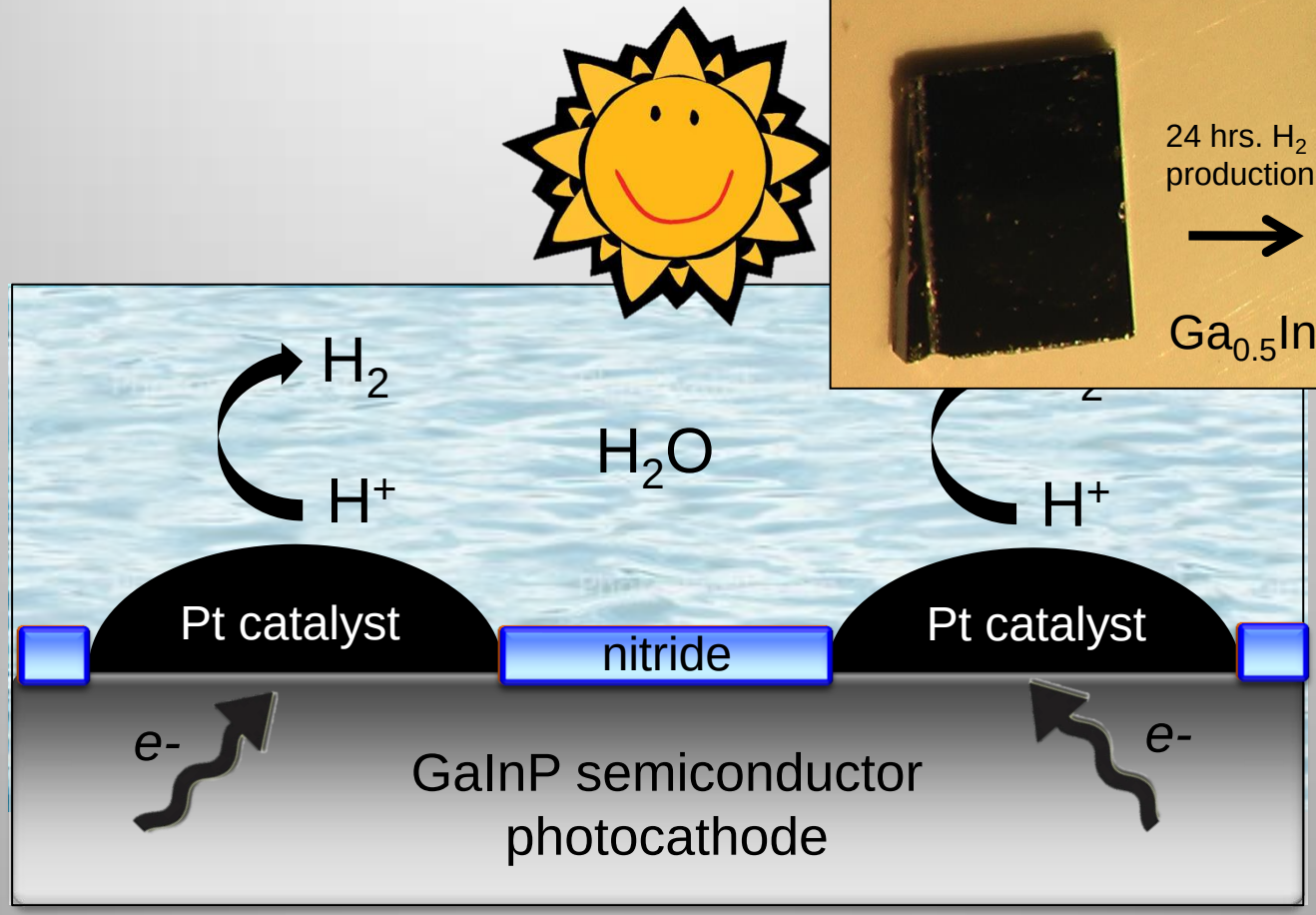
Solar energy storage and chemical fuel production

- **Option 1: Storage in batteries and capacitors**
  - Short-term storage
  - Short distances from generation source

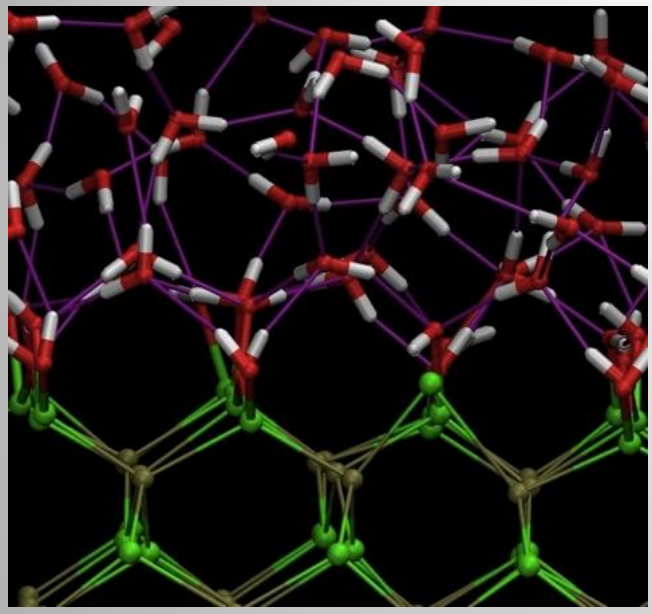


- **Option 2: Storage in chemical bonds**
  - Longer-term storage
  - Transportable over long distances
  - Produce usable fuel
  - Want a carbon-free (or closed-carbon) reaction cycle

Sun + Water + Photoelectrics = Hydrogen Fuel

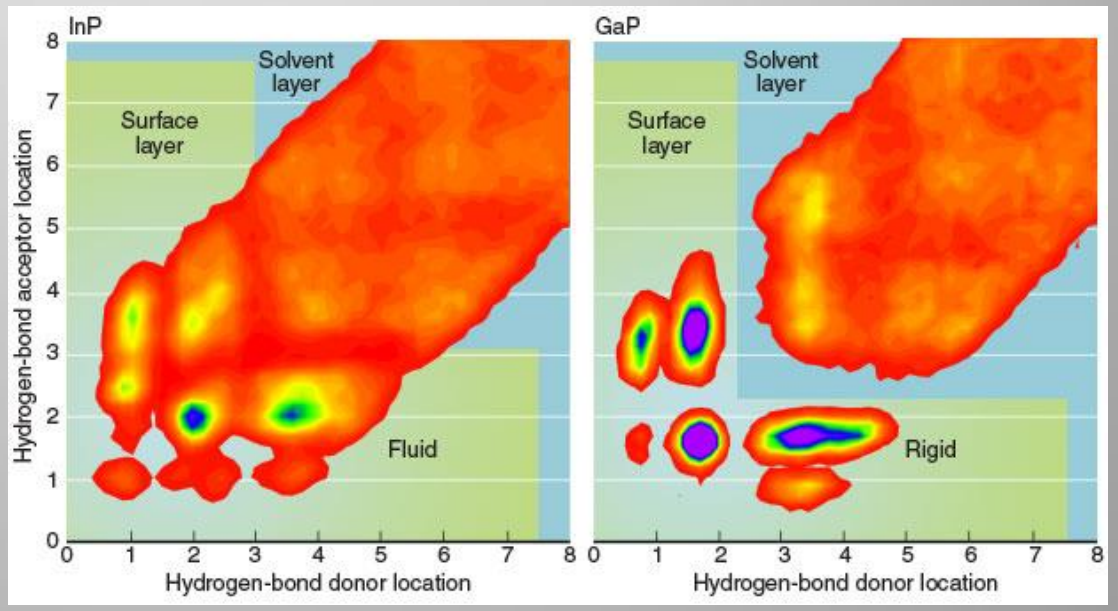


Insights from Molecular Dynamics

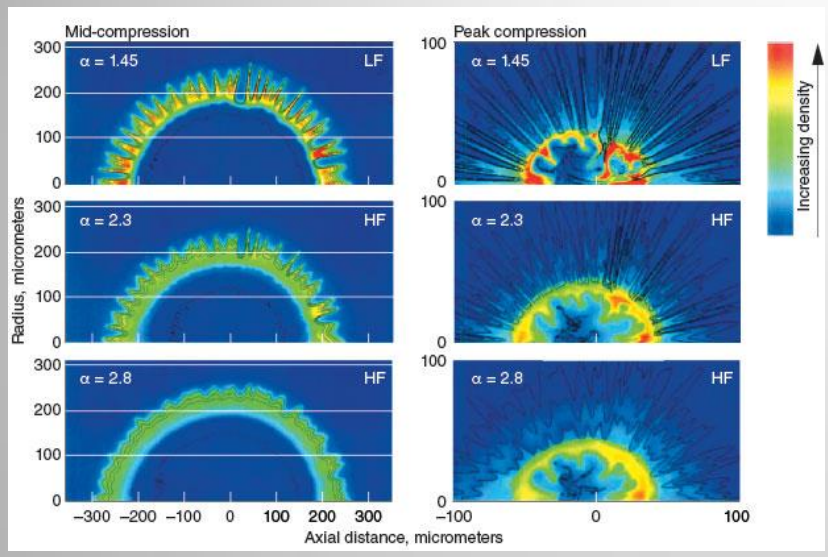


- Brandon Wood, Tadashi Ogitsu, Wooni Choi
- DOE's 2014 Hydrogen Production R&D Award, presented by EERE's Fuel Cell Technologies Office
- https://str.llnl.gov/april-2015/wood

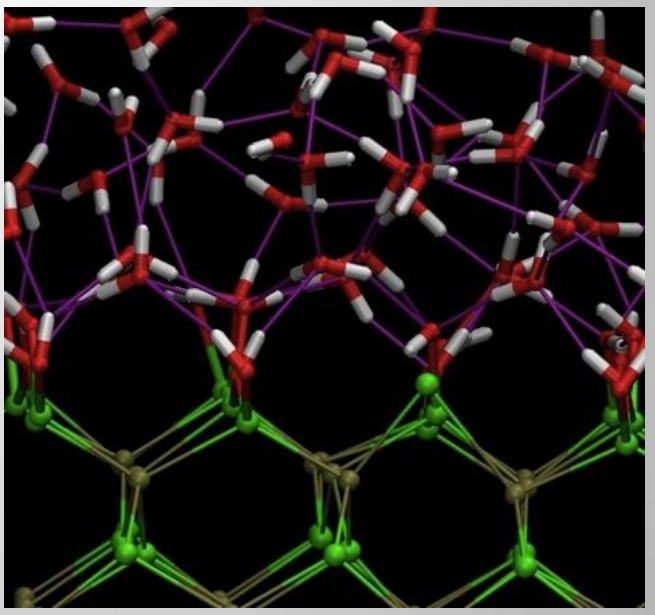
Simulations at semiconductor-water interface explain differences in efficiency and corrosion



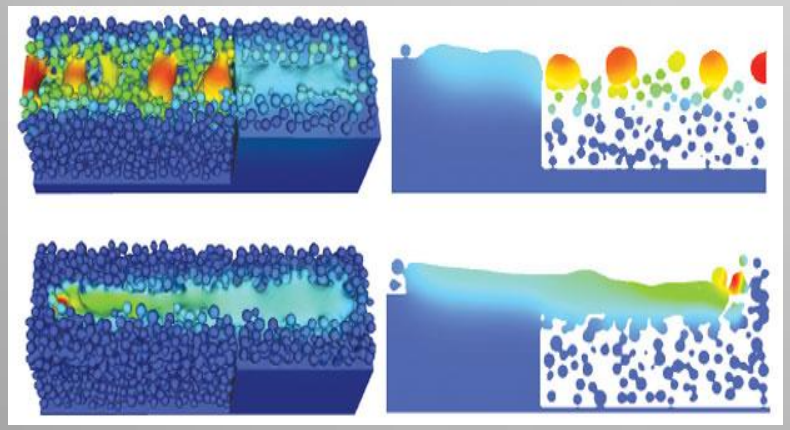
Science Recap: MVAPICH was there!



Fusion Science



Hydrogen Fuel



Additive Manufacturing

Future systems and challenges

Commodity Technology Systems

- New Linux clusters to replace existing systems
- Roughly same number of nodes, updated hardware (CPUs, memory, network)
- Delivery in 2016 through 2018
- Systems to last about 5 years
- Contract not finalized but MVAPICH stands good chance to continue as primary MPI
- Potentially many more years with MVAPICH

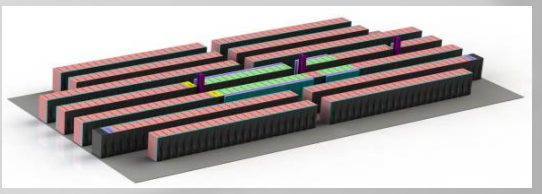
CORAL[LLNL] Sierra details

Compute System

2.1 – 2.7 PB Memory
120 -150 PFLOPS
10 MW

Compute Rack

Standard 19”
Warm water cooling



Compute Node

POWER® Architecture Processor
NVIDIA® Volta™
NVMe-compatible PCIe 800GB SSD
> 512 GB DDR4 + HBM
Coherent Shared Memory

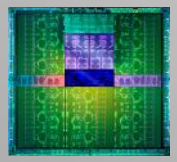
Components

IBM POWER

- NVLink™

NVIDIA Volta

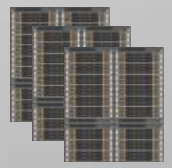
- HBM
- NVLink



Mellanox® Interconnect

Dual-rail EDR Infiniband®

OpenPOWER™



GPFS™ File System

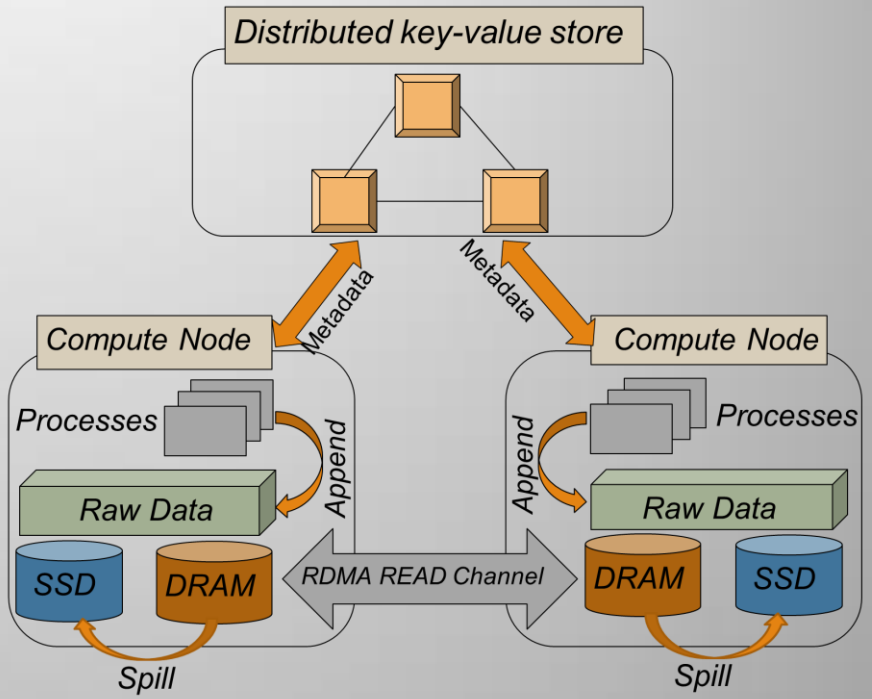
120 PB usable storage
1.2/1.0 TB/s R/W bandwidth

Challenge #1: Port to POWER processor

- MVAPICH team already experts with Mellanox and NVIDIA
- Just need port to POWER
- CORAL systems (Sierra and Summit) will be two of the world's fastest when they come online
- Let's get MVAPICH on these systems!

Challenge #2: Efficient MPI_THREAD_MULTIPLE and faster blocking communication

- Compute node storage driving requirements for async services
- Desire to use threads or second MPI job running in background of main app



BurstFS research prototype by Teng Wang

Challenge #3: Stabilize ABI to avoid recompiles

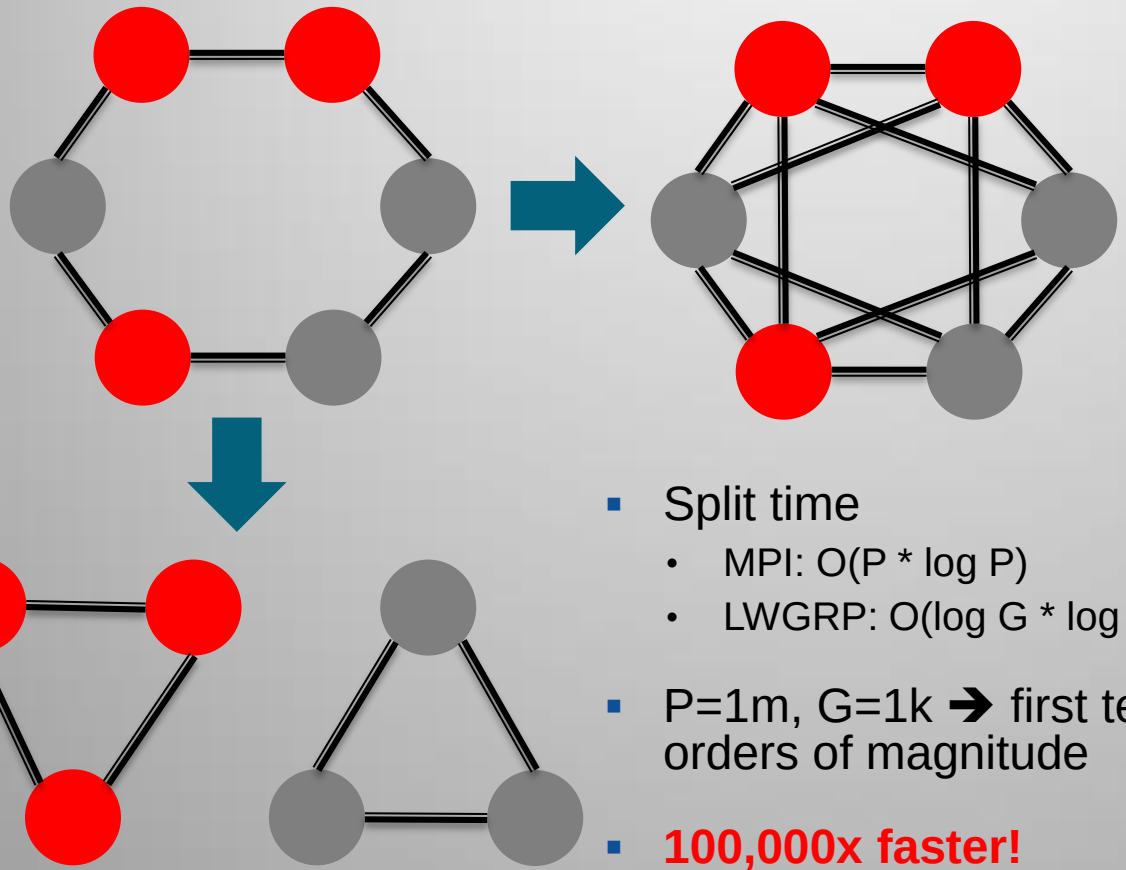
- Large LLNL codes takes 1 day to recompile app and all libraries
- Must then be tested and verified
- ABI breaks compatibility between versions
  - MV1-1.2 → MV2-1.7 → MV2-1.9 → MV2-2.1
  - Example solution: MPI-Adapter (Japan) or MorphMPI

Challenge #4: Improve MV2 performance

- Make MV2 faster than MV1
  - Some users still report a slowdown when switching to MV2
- Possible instruction bloat between the two?
  - e.g., use HPCToolkit and other profiling tools to find and remove bottlenecks and unnecessary code

Challenge #5: Fast MPI_Comm create and destroy

https://github.com/hpc/lwgrp



- Group representation
  - MPI: O(P)
  - LWGRP: O(log P)
- P=1m ➔ memory reduced 5 orders of magnitude
- **100,000x less memory!**
- log(P) collectives: Barrier, Bcast, Allreduce, Scan, Allgather, Alltoall, etc.

- Split time
  - MPI: O(P * log P)
  - LWGRP: O(log G * log P)
- P=1m, G=1k ➔ first term reduced 5 orders of magnitude
- **100,000x faster!**

Challenge Recap

- Port to IBM Power Processor
- Efficient MPI_THREAD_MULTIPLE and efficient blocking mode
- ABI compatibility between versions
- Make MV2 faster than MV1
- Find and eliminate every O(P) scaling term
  - Faster MPI_Comm create and destroy

Dear MVAPICH team,

Thank you !

And keep it up!

Sincerely,

Your friend, Science