Scalability of Next Generation Global NWP

John Michalakes
UCAR

Alex Reinecke, Kevin Viner, James Doyle
Naval Research Laboratory, Marine Meteorology Division

William Skamarock, Michael Duda
NCAR Mesoscale and Microscale Meteorology Division

Andreas Müller
Naval Postgraduate School

Tom Henderson
Spire Global Inc.

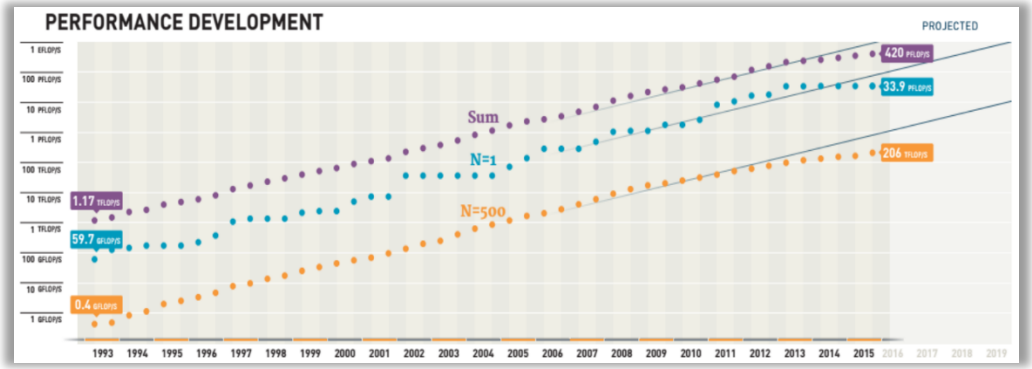
17th WRF Users Workshop
28 June 2016

UCAR NCAR

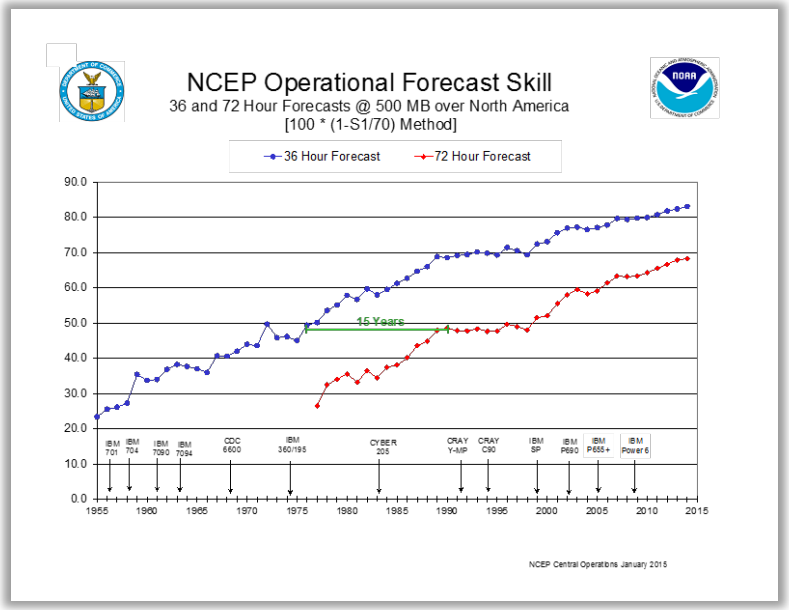
U.S. NAVAL RESEARCH LABORATORY

NWP and the HPC Imperative

- Exponential growth in HPC drives linear improvement in forecast skill
- All future gains for:
  - Higher resolution
  - Higher numerical order
  - More ensembles
  - Better physics
- Will mean adapting to new architectures that require reengineering for
  - More parallelism
  - Better locality (FLOPS/byte)
- New models, methods



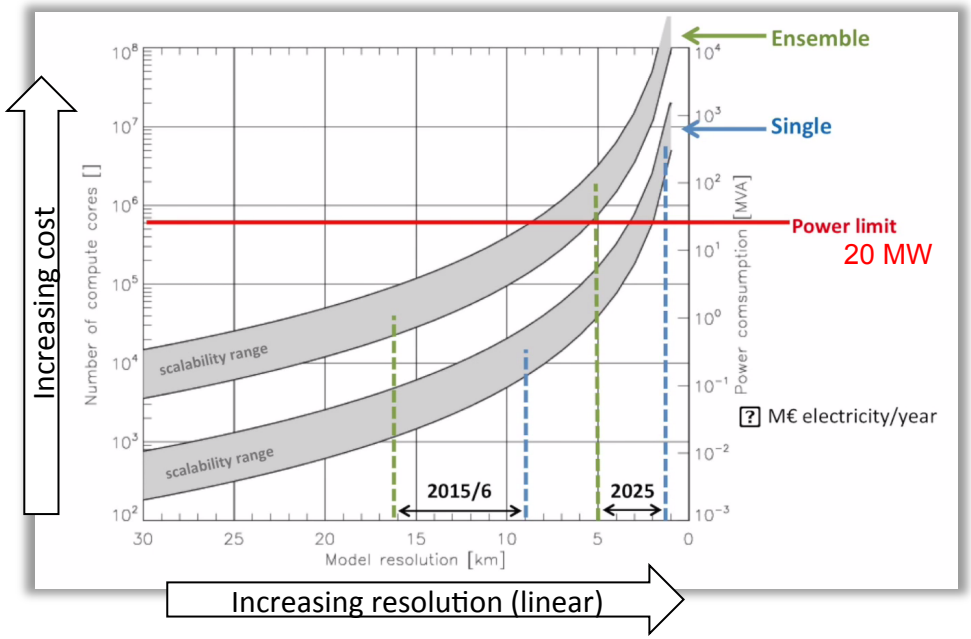
http://www.hpcwire.com/wp-content/uploads/2015/11/TOP500_201511_Poster.pdf



NCEP Central Operations, 2015

What are we scaling to? *Higher resolution*

- Benefits: improve forecast by resolving:
  - + Effects of complex terrain
  - + Dynamical features previously handled only as sub-grid
- Computational considerations:
  - – Uniform high (3 km and finer) resolution for operational NWP will be very costly, even assuming perfect parallel efficiency

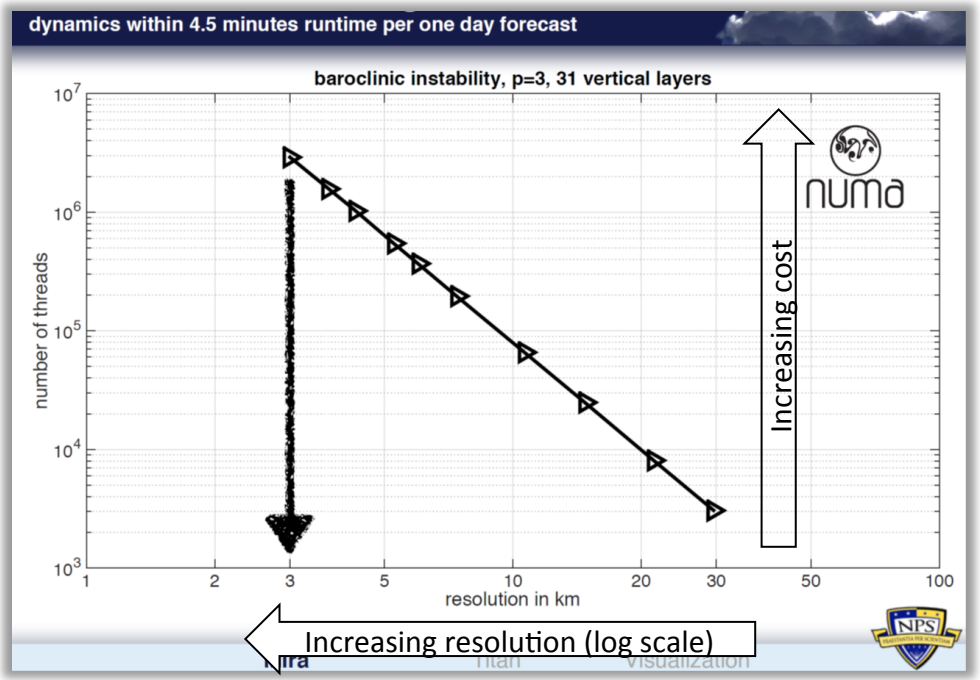


Bauer, Thorpe, and Brunet. "The quiet revolution of numerical weather prediction." *Nature* 525.7567 (2015): 47-55.

Notations by Erland Källén. Weather Prediction and the Scalability Challenge. Keynote presentation. Exascale Applications & Software Conference. April 2016. Stockholm

What are we scaling to? *Higher resolution*

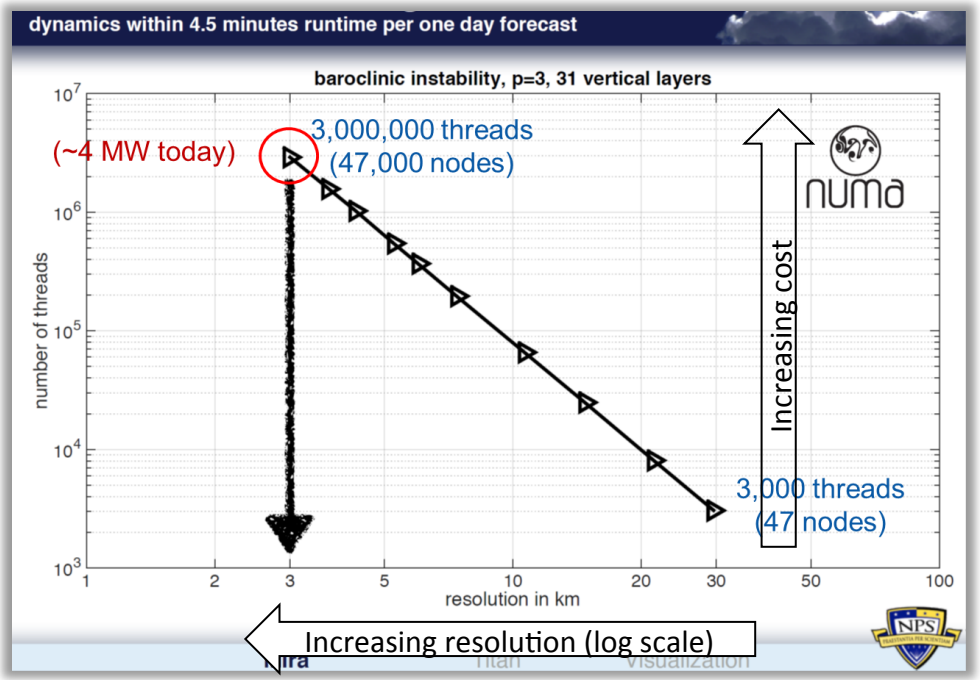
- Benefits: improve forecast by resolving:
  - + Effects of complex terrain
  - + Dynamical features previously handled only as sub-grid
- Computational considerations:
  - – Uniform high (3 km and finer) resolution for operational NWP will be very costly, even assuming perfect parallel efficiency



Müller, A., Kopera, M.A., Marras, S., Wilcox, L.C., Isaac, T., and Giraldo F.X. Strong Scaling for Numerical Weather Prediction at Petascale with the Atmospheric Model NUMA. 2016. Submitted http://arxiv.org/abs/1511.01561

What are we scaling to? *Higher resolution*

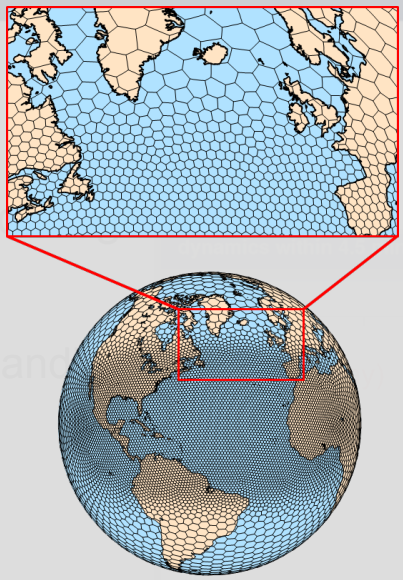
- Benefits: improve forecast by resolving:
  - + Effects of complex terrain
  - + Dynamical features previously handled only as sub-grid
- Computational considerations:
  - – Uniform high (3 km and finer) resolution for operational NWP will be very costly, even assuming perfect parallel efficiency



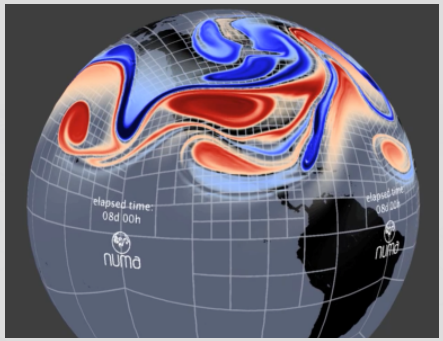
Müller, A., Kopera, M.A., Marras, S., Wilcox, L.C., Isaac, T., and Giraldo F.X. Strong Scaling for Numerical Weather Prediction at Petascale with the Atmospheric Model NUMA. 2016. Submitted http://arxiv.org/abs/1511.01561

What are we scaling to? *Higher resolution*

- Benefits: improve forecast by
  - + Effects of complex terrain
  - + Dynamical features previous only as sub-grid
- Computational considerations:
  - – Uniform high (3 km and fine) resolution for operational NWP may not be feasible, even assuming perfect parallel efficiency
  - – **Non-uniform resolution (nesting, mesh refinement, grid stretching) needed to focus costly resolution**



MPAS



NUMA
Courtesy: Frank Giraldo
NPS

Müller, A., Kopera, M.A., Marras, S., Wilcox, L.C., Isaac, T., and Giraldo F.X. Strong Scaling for Numerical Weather Prediction at Petascale with the Atmospheric Model NUMA. 2016. Submitted http://arxiv.org/abs/1511.01561

What are we scaling to? *Higher order*

- More expensive but cost-effective if higher accuracy is beneficial.
  - – Errors in physics and initial conditions dominate
  - \+ Resolve larger range of waves with lower dissipation
  - \+ For adaptive grids, better transitioning between grid resolutions
  - \+ Even with physics, higher-order numerics can reduce dynamical biases

Mavriplis, Catherine. **The Challenges of High Order Methods in Numerical Weather Prediction** Lecture Notes in Computational Science and Engineering. Vol. 76. 17 Sept. 2010. Springer. pp 255-266

Reinecke, A. Patrick A., and Dale Durran. **The overamplification of gravity waves in numerical solutions to flow over topography.** MWR 137.5 (2009): 1533-1549.

- Computational considerations for next generation HPC
  - – More expensive in terms of operations
  - \+ Good locality, computational intensity,
  - \+ Less communication, better interprocessor scaling

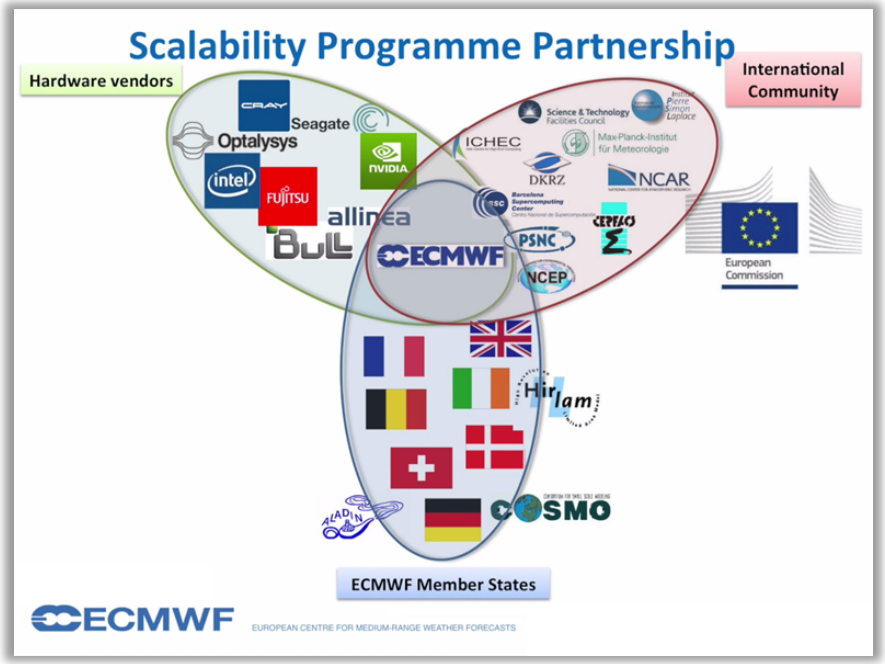
Alternative view: use *less* accurate methods and "reinvest" savings back into higher resolution, more ensemble members, etc.

- – Düben and Palmer (U. Oxford) suggest reducing floating point precision and tolerating some level of processor errors that result from overclocking

Düben, Peter D., and T. N. Palmer. **Benchmark tests for numerical weather forecasts on inexact hardware.** Monthly Weather Review 142.10 (2014): 3809-3829.

Current efforts

- ECMWF Scalability Program
  - Energy-efficient Scalable Algorithms for Weather Prediction at Exascale (ESCAPE) www.hpc-escape.eu
  - Panta Rhei: *http://www.ecmwf.int/en/research/projects/pantarhei*
- UKMO
  - Gung Ho/LFRic https://puma.nerc.ac.uk/trac/GungHo
- NOAA
  - NGGPS (Next Generation Global Prediction System)
  - MPAS and NEPTUNE

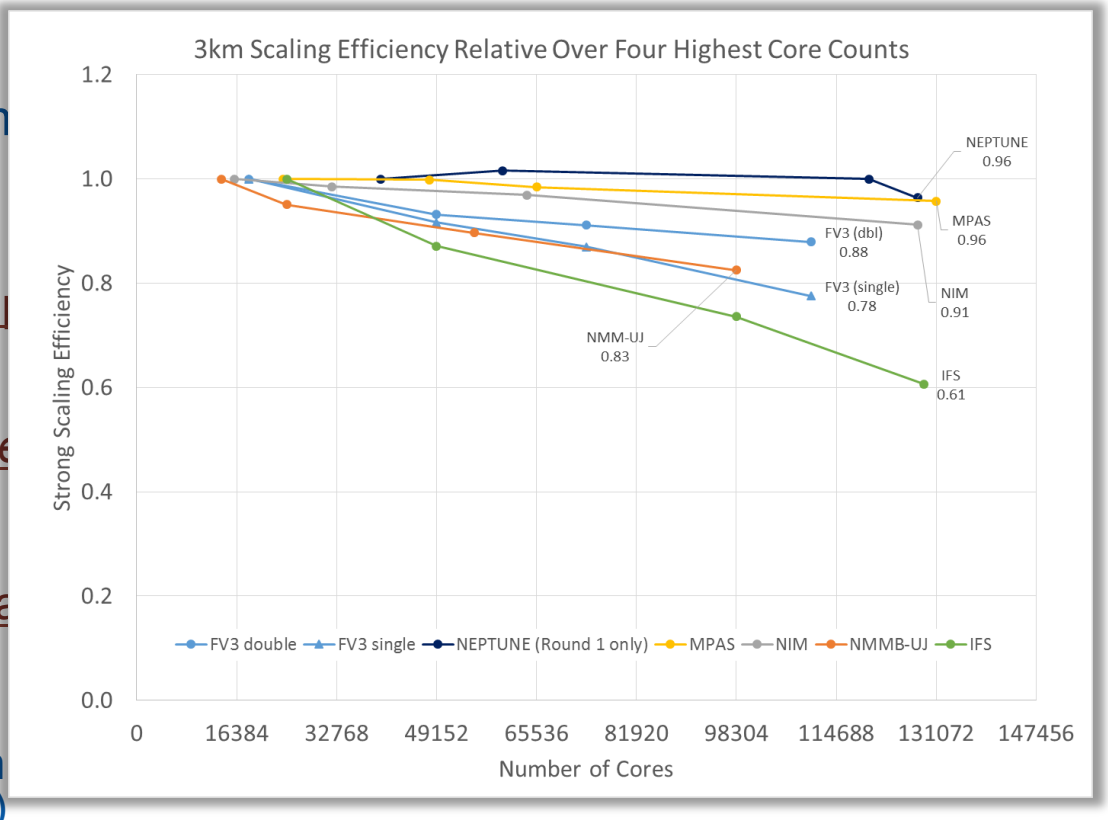


Erland Källén. **Weather Prediction and the Scalability Challenge.** Keynote presentation. Exascale Applications & Software Conference. April 2016. Stockholm

Current efforts

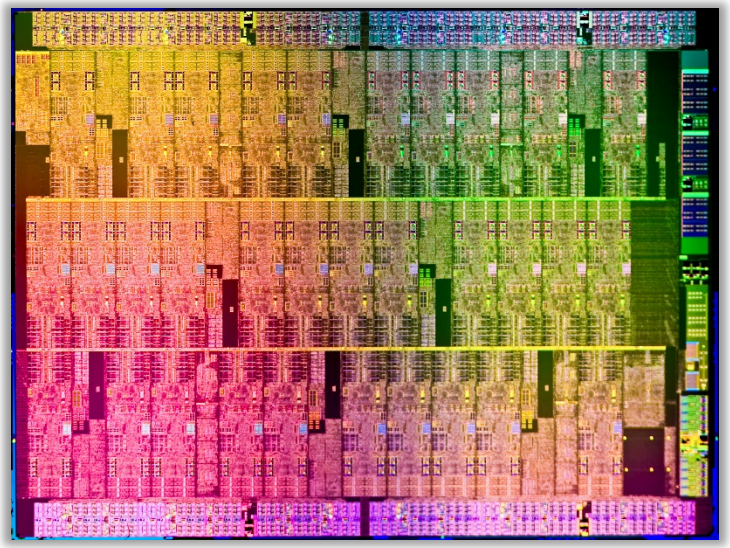
- ECMWF Scalability Program
  - Energy-efficient Scalable Algorithms for Weather Prediction at Exascale (ESCAPE) www.hpc-escap
  - Panta Rhei: *http://www.ecmwf.int/en/research/projects/pantarhe*
- UKMO
  - Gung Ho/LFRic https://puma.nerc.ac.uk/tra GungHo
- NOAA
  - NGGPS (Next Generation Global Prediction System)
  - MPAS and NEPTUNE

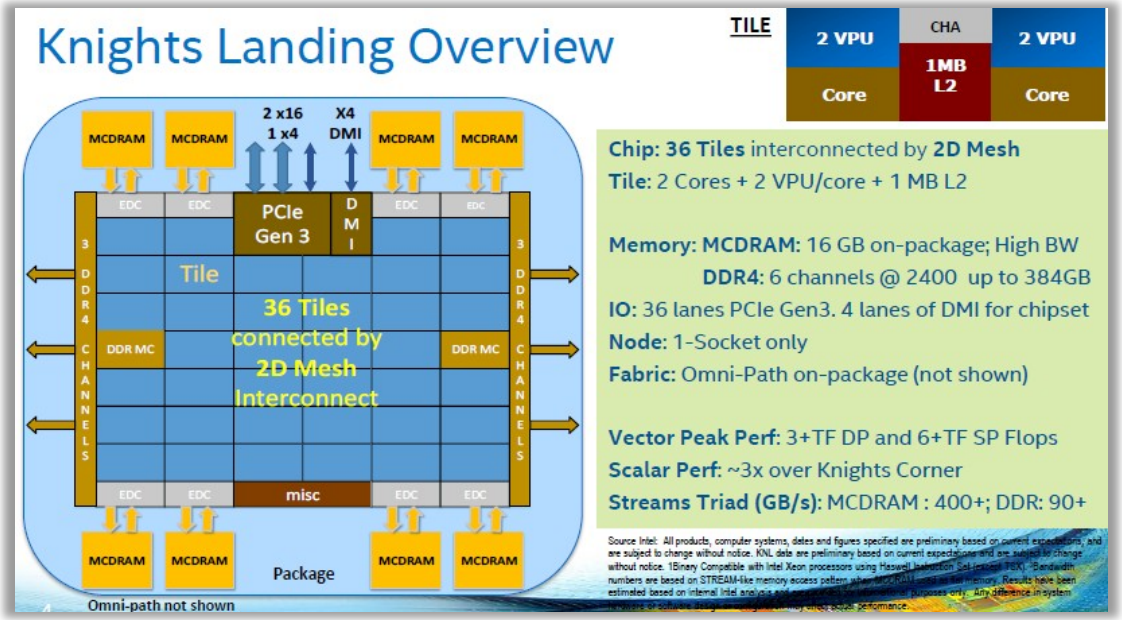
(Higher is Better)



NGGPS Phase-1 benchmarking report of the Advanced Computing Evaluation Committee. April, 2015 http://www.nws.noaa.gov/ost/nggps/dycoretesting.html

Hardware: Knights Landing



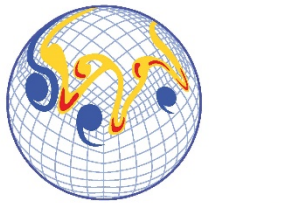


- Intel Xeon Phi 7250 (Knights Landing) announced at ISC'16 this month
  - 14 nanometer feature size, > 8 billion transistors
  - 68 cores, 1.4 GHz modified "Silvermont" with out-of-order instruction execution
  - Two 512-bit wide Vector Processing Units per core
  - Peak ~3 TF/s double precision, ~6 TF/s single precision
  - 16 GB MCDRAM (on-chip) memory, > 400 GB/s bandwidth
  - "Hostless" – no separate host processor and no "offload" programming
  - Binary compatible ISA (with extensions for AVX-512 vector instructions

Models: NEPTUNE/NUMA

numa

- **NUMA** = **N**on-hydrostatic **U**nified **M**odel of the **A**tmosphere
- engine inside the Navy's next generation weather prediction system NEPTUNE (**N**avy's **E**nvironment **P**rediction sys**T**em **U**sing the **N**uma **E**ngine)
- developed by **Prof. Francis X. Giraldo** and generations of postdocs

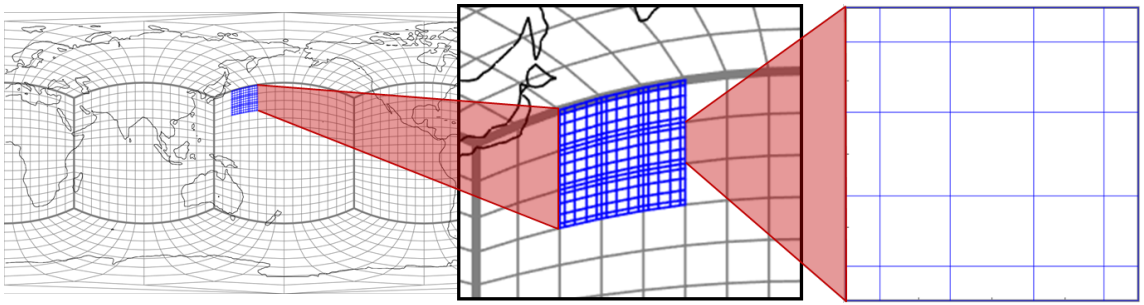


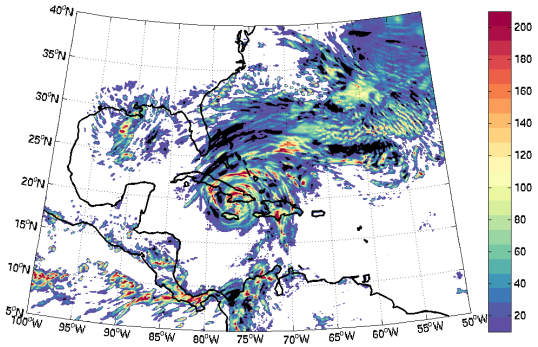
numa

Andreas Mueller, NPS, Monterey, CA; and M. Kopera, S. Marras, and F. X. Giraldo. "Towards Operational Weather Prediction at 3.0km Global Resolution With the Dynamical Core NUMA". 96th Amer. Met. Society Annual Mtg. January, 2016. https://ams.confex.com/ams/96Annual/webprogram/Paper288883.html

Models: NEPTUNE/NUMA

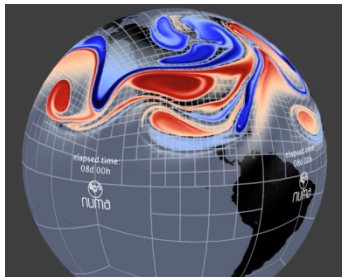
- Spectral element
  - 4th, 5th and higher-order* continuous Galerkin (discontinuous planned)
  - Cubed Sphere (also icosahedral)
  - Adaptive Mesh Refinement (AMR) in development





NEPTUNE 72-h forecast (5 km resolution) of accumulated precipitation for Hurr. Sandy

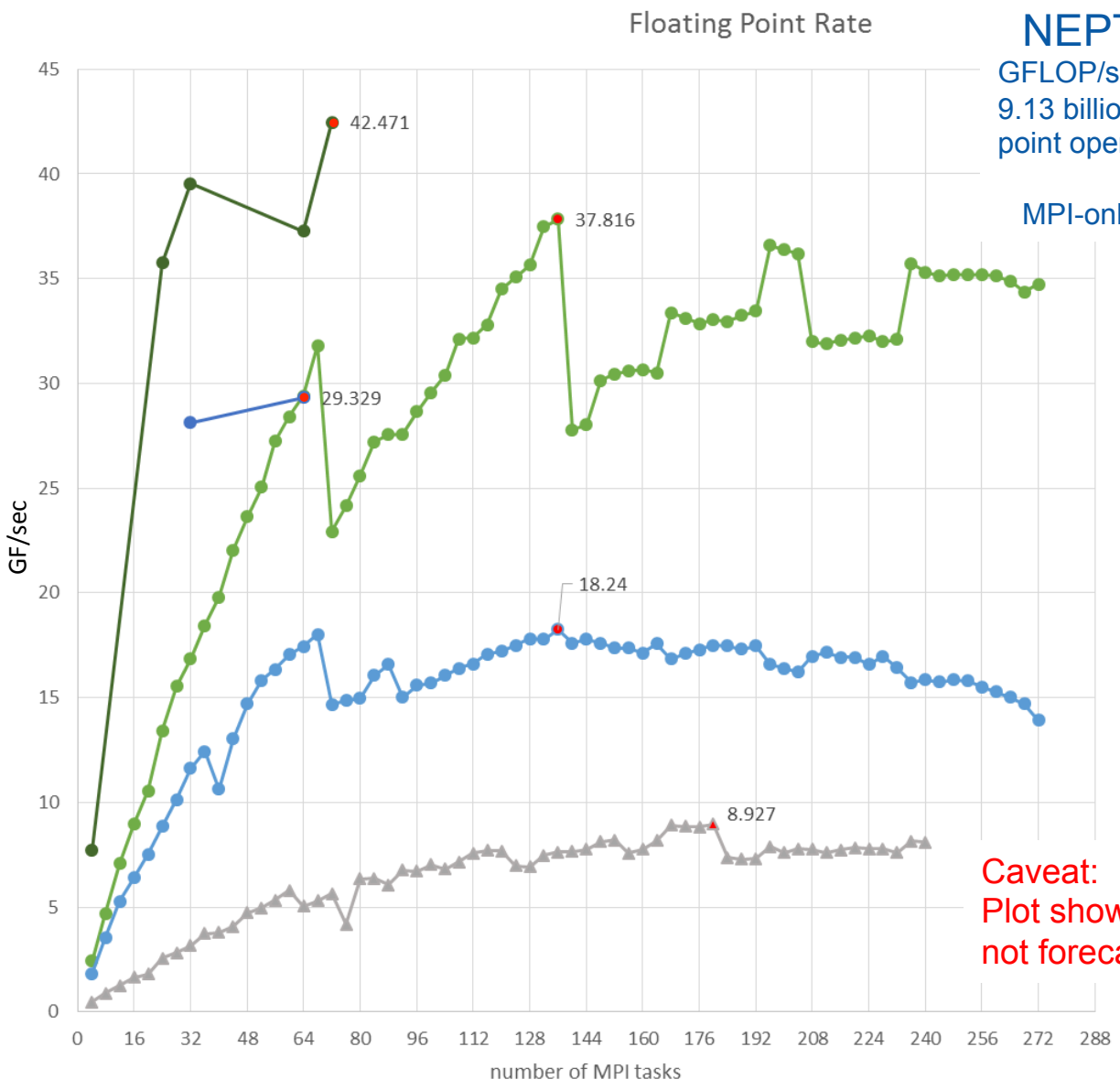
- Computationally dense but highly scalable
  - Constant width-one halo communication
  - Good locality for next generation HPC



Example of Adaptive Grid tracking a severe event courtesy: Frank Giraldo, NPS

*"This is not the same 'order' as is used to identify the leading term of the error in finite-difference schemes, which in fact describes accuracy. Evaluation of Gaussian quadrature over N+1 LGL quadrature points will be exact to machine precision as long as the polynomial integrand is of the order 2x(N-1) -3, or less." Gabersek et al. MWR Apr 2012.DOI: 10.1175/MWR-D-11-00144.1

NEPTUNE Performance on 1 Node Knights Landing



NEPTUNE

GFLOP/second, based on count of 9.13 billion *double precision* floating point operations per time step

MPI-only

Caveat:
Plot shows floating point rate, not forecast speed.

NEPTUNE Performance on 1 Node Knights Landing

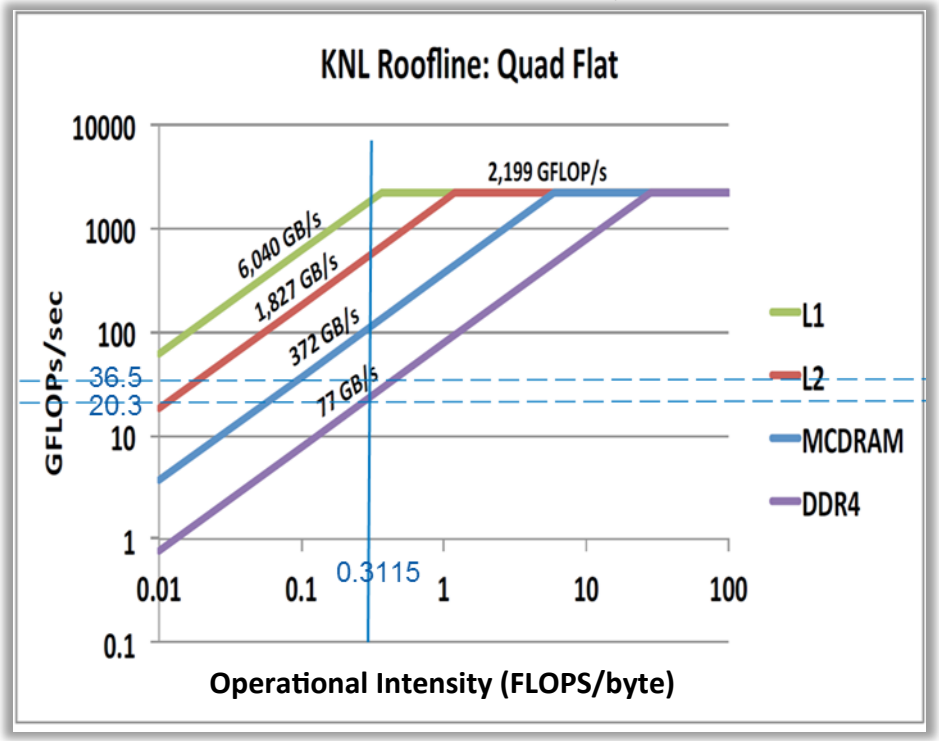
	4rd order NEPTUNE E14P3L40	5th order NEPTUNE E10P4L41	P3/P4
resolution	2	2	1.00
nelem	15288	6000	2.55
npoints (1 task)	423440	393682	1.08
npoints (136 tasks)	562960	521766	1.08
dt	2	1	2.00
Op count (SDE) (one step)	9,131,287,622	8,136,765,866	1.12
ops per cell-step (1 task)	21565	20668	1.04
Bytes r/w reg-L1 (SDE)	95,941,719,268	81,973,718,952	1.17
Bytes r/w L2-MCDRAM (VTune)	29,318,220,928	27,278,494,016	1.07
Arithmetic Intensity	0.0952	0.0993	0.96
Operational Intensity	0.3115	0.2983	1.04
MCDRAM Time (1 step) (sec)	0.250143	0.234722	1.07
MCDRAM GF/s	**36.504**	34.665	1.05
MCDRAM GB/s	1.831	1.815	1.01

- Performance model in development
  - Explain observed performance
  - Identify bottlenecks for improvement
  - Predict performance on new platforms
- Factors
  - Resolution, order and time step
  - FP operations and data traffic
  - Operational intensity
  - Machine characteristics (for Roofline)
- Early results (left)
  - Operations, intensity and floating point rate *decrease* with higher order

NEPTUNE Performance on 1 Node Knights Landing

	4rd order NEPTUNE E14P3L40	5th order NEPTUNE E10P4L41	P3/P4
resolution	2	2	1.00
nelem	15288	6000	2.55
npoints (1 task)	423440	393682	1.08
npoints (136 tasks)	562960	521766	1.08
dt	2	1	2.00
Op count (SDE) (one step)	9,131,287,622	8,136,765,866	1.12
ops per cell-step (1 task)	21565	20668	1.04
Bytes r/w reg-L1 (SDE)	95,941,719,268	81,973,718,952	1.17
Bytes r/w L2-MCDRAM (VTune)	29,318,220,928	27,278,494,016	1.07
Arithmetic Intensity	0.0952	0.0993	0.96
Operational Intensity	0.3115	0.2983	1.04
MCDRAM Time (1 step) (sec)	0.250143	0.234722	1.07
MCDRAM GF/s	**36.504**	34.665	1.05
MCDRAM GB/s	1.831	1.815	1.01

D. Doerfler et al. **Applying the Roofline Performance Model to the Intel Xeon Phi Knights Landing Processor.** IXPUG Workshop ISC 2016, June 23rd, 2016 Frankfurt, Germany

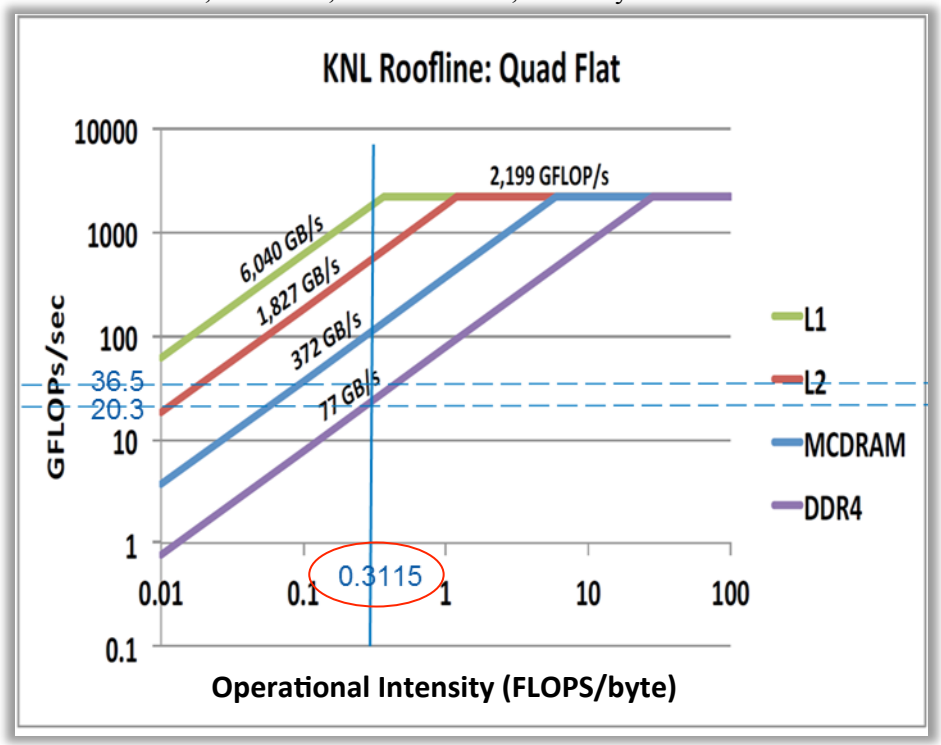


Operational intensity of 0.31 flops/byte is bandwidth-limited from KNL DRAM but has bandwidth to spare from MCDRAM.
With vector and other improvements performance should pass 100 GF/s according to Roofline model.
Improvements to Operational Intensity will raise ceiling further.

NEPTUNE Performance on 1 Node Knights Landing

	4rd order NEPTUNE E14P3L40	5th order NEPTUNE E10P4L41	P3/P4
resolution	2	2	1.00
nelem	15288	6000	2.55
npoints (1 task)	423440	393682	1.08
npoints (136 tasks)	562960	521766	1.08
dt	2	1	2.00
Op count (SDE) (one step)	9,131,287,622	8,136,765,866	1.12
ops per cell-step (1 task)	21565	20668	1.04
Bytes r/w reg-L1 (SDE)	95,941,719,268	81,973,718,952	1.17
Bytes r/w L2-MCDRAM (VTune)	29,318,220,928	27,278,494,016	1.07
Arithmetic Intensity	0.0952	0.0993	0.96
Operational Intensity	0.3115	0.2983	1.04
MCDRAM Time (1 step) (sec)	0.250143	0.234722	1.07
MCDRAM GF/s	**36.504**	34.665	1.05
MCDRAM GB/s	1.831	1.815	1.01

D. Doerfler et al. **Applying the Roofline Performance Model to the Intel Xeon Phi Knights Landing Processor.** IXPUG Workshop ISC 2016, June 23rd, 2016 Frankfurt, Germany

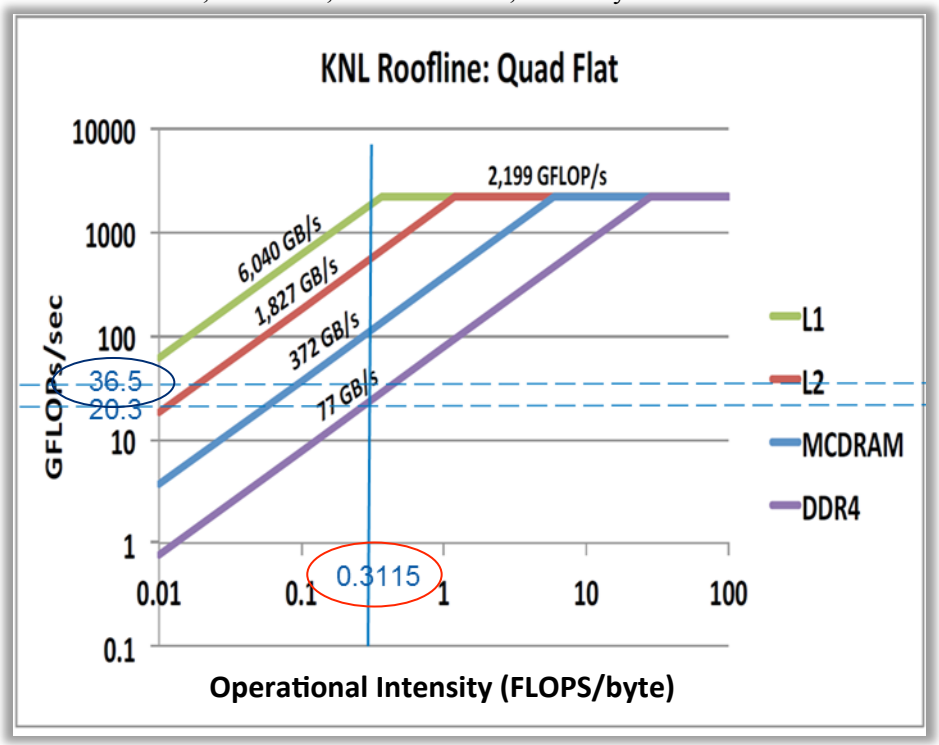


Operational intensity of 0.31 flops/byte is bandwidth-limited from KNL DRAM but has bandwidth to spare from MCDRAM.
With vector and other improvements performance should pass 100 GF/s according to Roofline model.
Improvements to Operational Intensity will raise ceiling further.

NEPTUNE Performance on 1 Node Knights Landing

	4rd order NEPTUNE E14P3L40	5th order NEPTUNE E10P4L41	P3/P4
resolution	2	2	1.00
nelem	15288	6000	2.55
npoints (1 task)	423440	393682	1.08
npoints (136 tasks)	562960	521766	1.08
dt	2	1	2.00
Op count (SDE) (one step)	9,131,287,622	8,136,765,866	1.12
ops per cell-step (1 task)	21565	20668	1.04
Bytes r/w reg-L1 (SDE)	95,941,719,268	81,973,718,952	1.17
Bytes r/w L2-MCDRAM (VTune)	29,318,220,928	27,278,494,016	1.07
Arithmetic Intensity	0.0952	0.0993	0.96
Operational Intensity	0.3115	0.2983	1.04
MCDRAM Time (1 step) (sec)	0.250143	0.234722	1.07
MCDRAM GF/s	**36.504**	34.665	1.05
MCDRAM GB/s	1.831	1.815	1.01

D. Doerfler et al. **Applying the Roofline Performance Model to the Intel Xeon Phi Knights Landing Processor.** IXPUG Workshop ISC 2016, June 23rd, 2016 Frankfurt, Germany

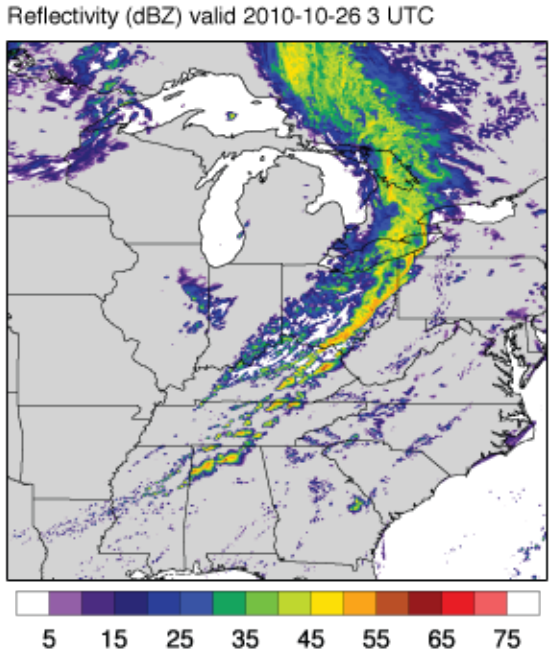


Operational intensity of 0.31 flops/byte is bandwidth-limited from KNL DRAM but has bandwidth to spare from MCDRAM.
With vector and other improvements performance should pass 100 GF/s according to Roofline model.
Improvements to Operational Intensity will raise ceiling further.

MPAS Model for Prediction Across Scales

Collaborative project between NCAR and LANL for developing atmosphere, ocean and other earth-system simulation components for use in climate, regional climate and weather studies

- Applications include global NWP and global atmospheric chemistry, regional climate, tropical cyclones, convection-permitting hazardous weather forecasting
- Finite Volume, C-grid
- Refinement capability
  - Centroidal Voronoi-tessellated unstructured mesh, allows arbitrary **in-place horizontal mesh** refinement
- HPC Readiness
  - Current release (v4.0) supports parallelism via MPI (horiz. domain decomp.)
  - Hybrid (MPI+OpenMP) parallelism implemented, undergoing testing



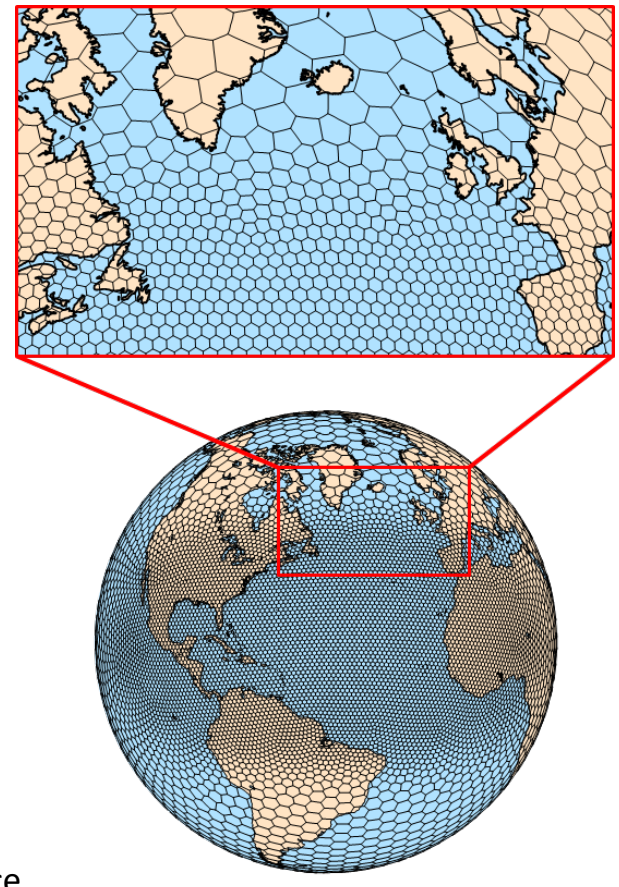
Simulated reflectivity diagnosed from the WSM6 hydrometeor fields in an MPAS 3-km global forecast initialized on 2010-10-23 at 00 UTC. Isolated severe convection is evident ahead of the cold front in agreement with observation.

Primarily funding: National Science Foundation and the DOE Office of Science

MPAS Model for Prediction Across Scales

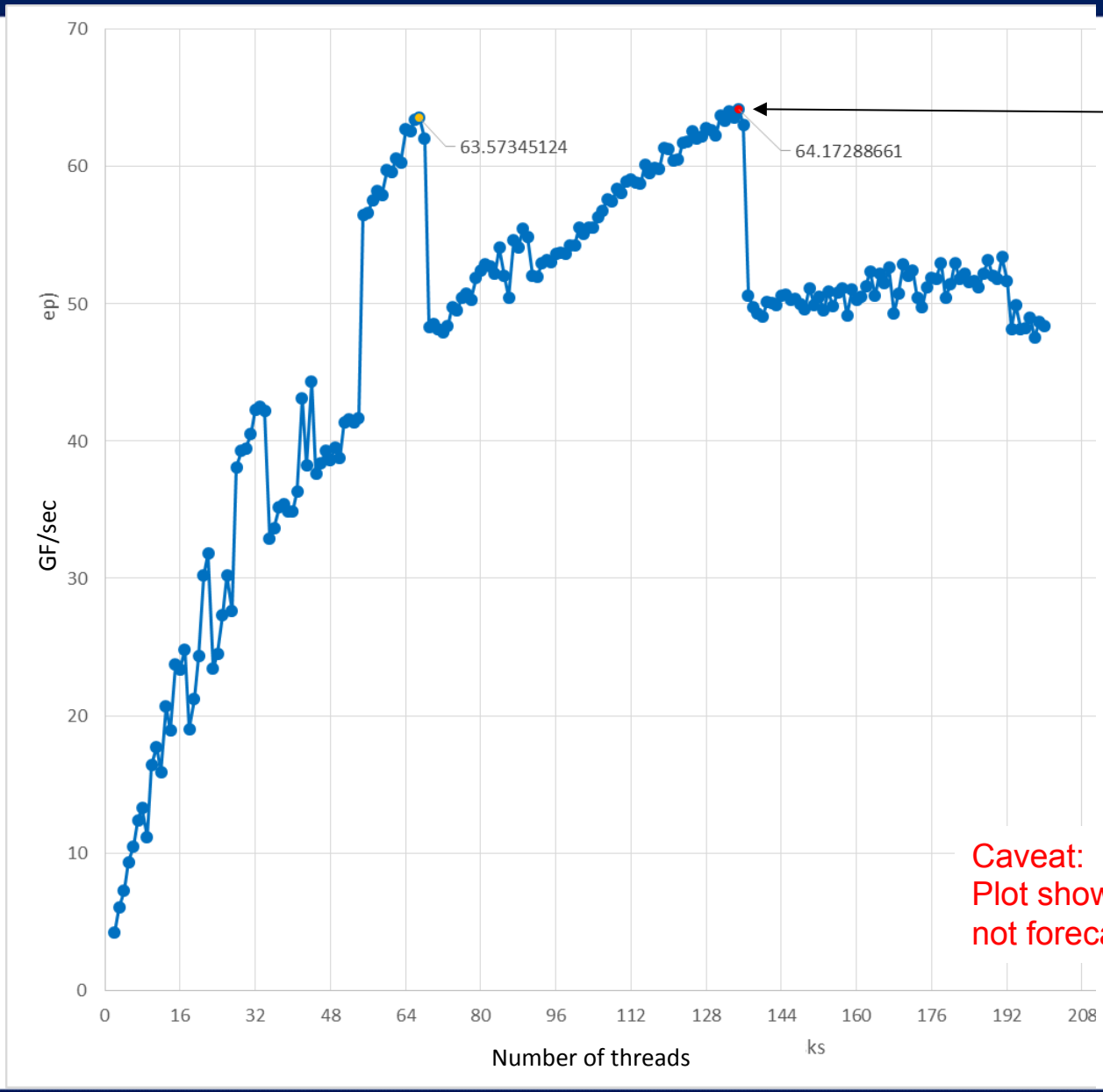
Collaborative project between NCAR and LANL for developing atmosphere, ocean and other earth-system simulation components for use in climate, regional climate and weather studies

- Applications include global NWP and global atmospheric chemistry, regional climate, tropical cyclones, convection-permitting hazardous weather forecasting
- Finite Volume, C-grid
- Refinement capability
  - Centroidal Voronoi-tessellated unstructured mesh, allows arbitrary **in-place horizontal mesh** refinement
- HPC Readiness
  - Current release (v4.0) supports parallelism via MPI (horiz. domain decomp.)
  - Hybrid (MPI+OpenMP) parallelism implemented, undergoing testing



Primarily funding: National Science Foundation and the DOE Office of Science

MPAS Performance on 1 Node Knights Landing



MPAS

GFLOP/second, based on count of 3.7 billion *single precision* floating point operations per dynamics step

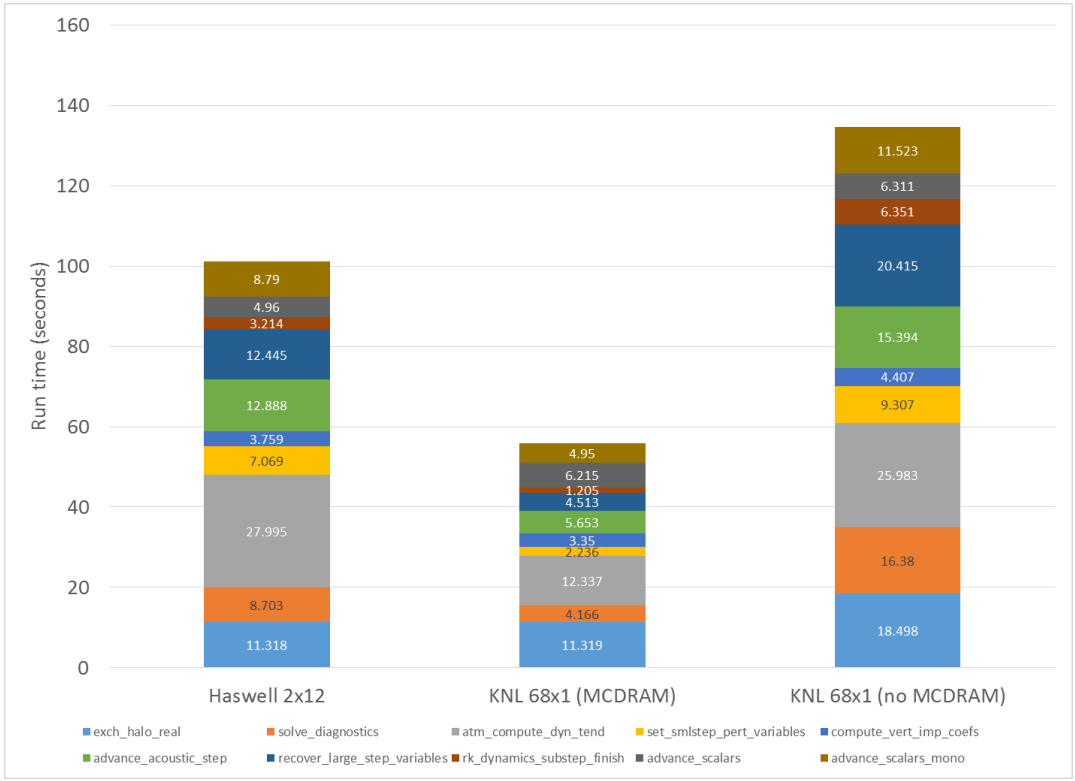
MPI and OpenMP
but best time shown here was single-threaded

KNL 7250 B0 68c (MCDRAM)

Caveat:
Plot shows floating point rate, not forecast speed.

NEPTUNE Performance on 1 Node Knights Landing

	MPAS Supercell
resolution	2 km
npoints	409680
dt (dynamics)	6
Op count (SDE) (one dyn step)	3,675,181,216
ops per cell-step (1 task)	8,971
Bytes r/w reg-L1 (SDE)	18,797,394,348
Bytes r/w L2-DRAM (VTune)	10,759,238,816
Bytes r/w L2-MCDRAM (VTune)	9,533,493,568
Arithmetic Intensity	0.196
Operational Intensity	0.386
DRAM Time (1 dyn step) (sec)	0.075
DRAM GF/s	49.2
MCDRAM Time (1 dyn step) (sec)	0.057
MCDRAM GF/s	64.2
MCDRAM peak GB/s	



MPAS Cost Breakdown on
Haswell and Knights Landing
Jim Rosinski, Tom Henderson, Mark Govett
NOAA/ESRL June 2016

Govett, Henderson and Rosinski found
Hybrid was slightly better than straight MPI
for MPAS running on Haswell (12x2 vs 2x12)

Summary: Scaling NWP to Future HPC

- New models such as MPAS and NUMA/NEPTUNE
  - Higher order, variable resolution
  - Demonstrated scalability
  - Focus is on computational efficiency to run *faster*
- Progress with accelerators
  - Early Intel Knights Landing results showing 1-2% of peak
  - Good news: with new MCDRAM, we aren't memory bandwidth bound
  - Need to improve vector utilization, other bottlenecks to reach > 10% of peak
  - Naval Postgraduate School's results speeding up NUMA show promise
- Ongoing efforts
  - Measurement, characterization and optimization of floating point and data movement
  - Establish and strengthen collaborations towards scalability for NWP