NOAA Operational Forecasting and the HPC Imperative

John Michalakes

NOAA/NCEP/Environmental Modeling Center
(IM Systems Group)
University of Colorado at Boulder

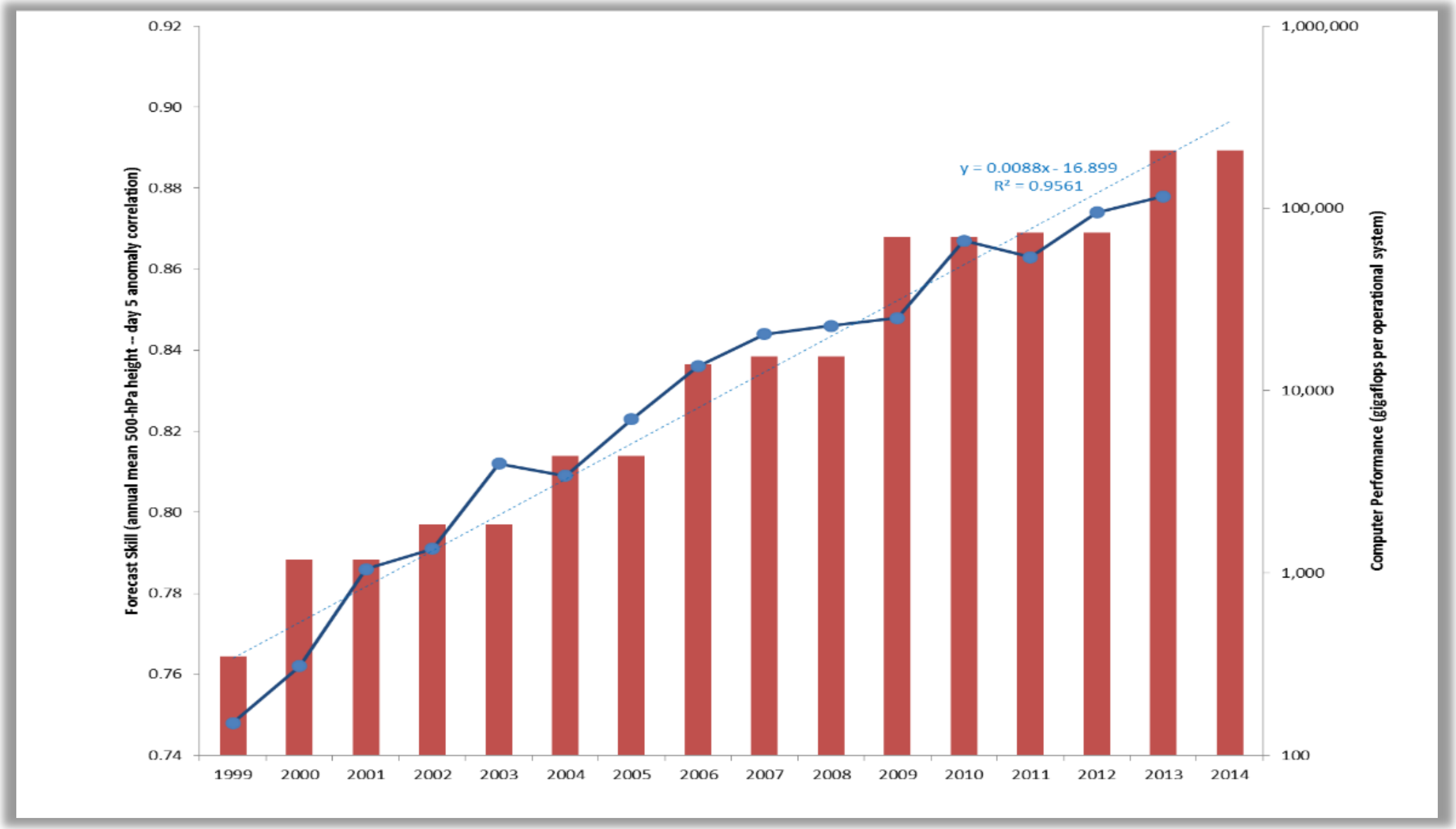
16th ECMWF Workshop on HPC in Meteorology
28 October 2014

Outline

- The HPC Imperative
- Next-Generation Global Prediction System
- Accelerators and NWP

HPC Imperative

- NWP is one of the first HPC applications and, with climate, has been one its key drivers
  - ➢ Exponential growth in HPC capability has translated directly to better forecasts and steadily improving value to the public

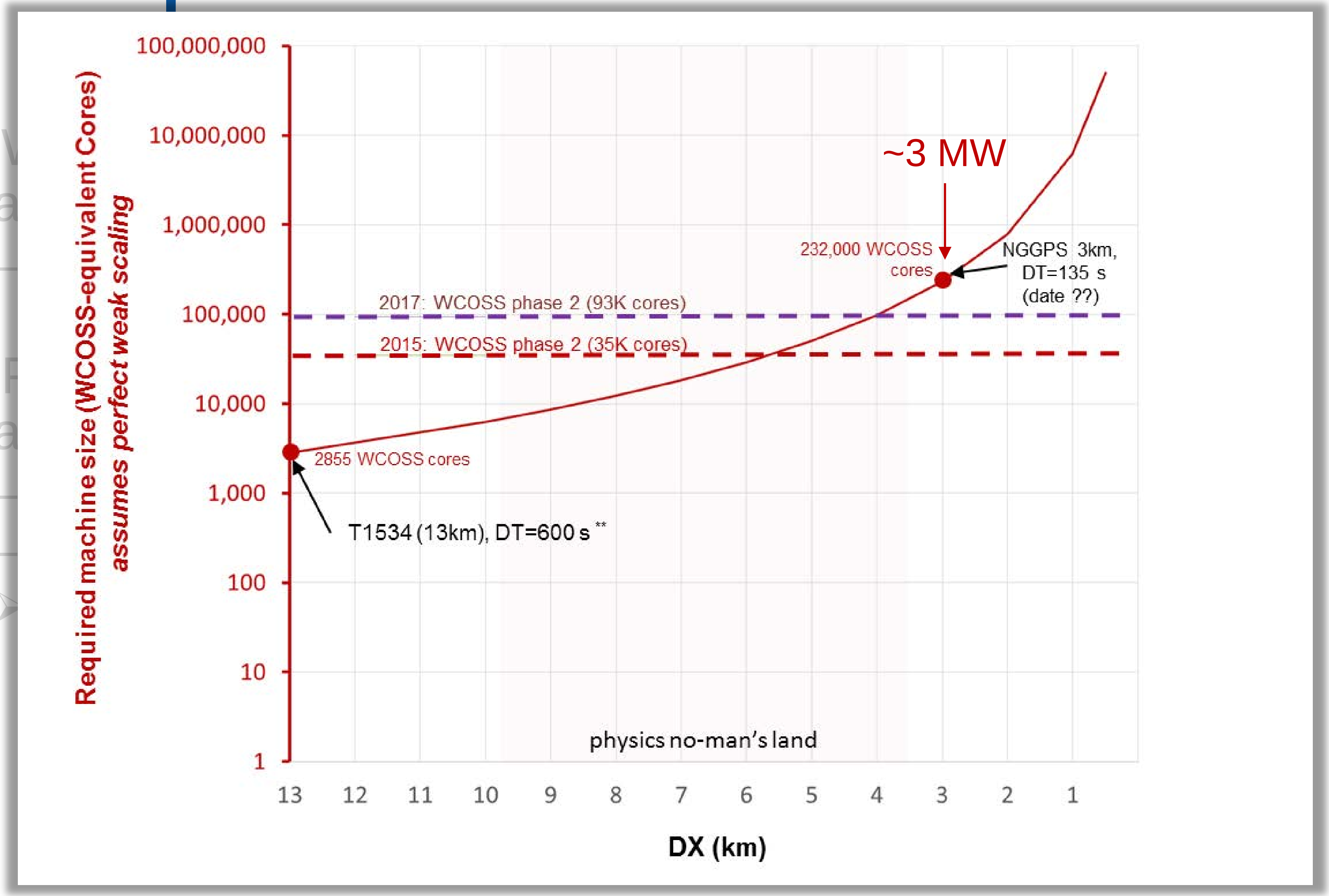


Fred Toepfer and Ed Mifflin - HFIP

HPC Imperative

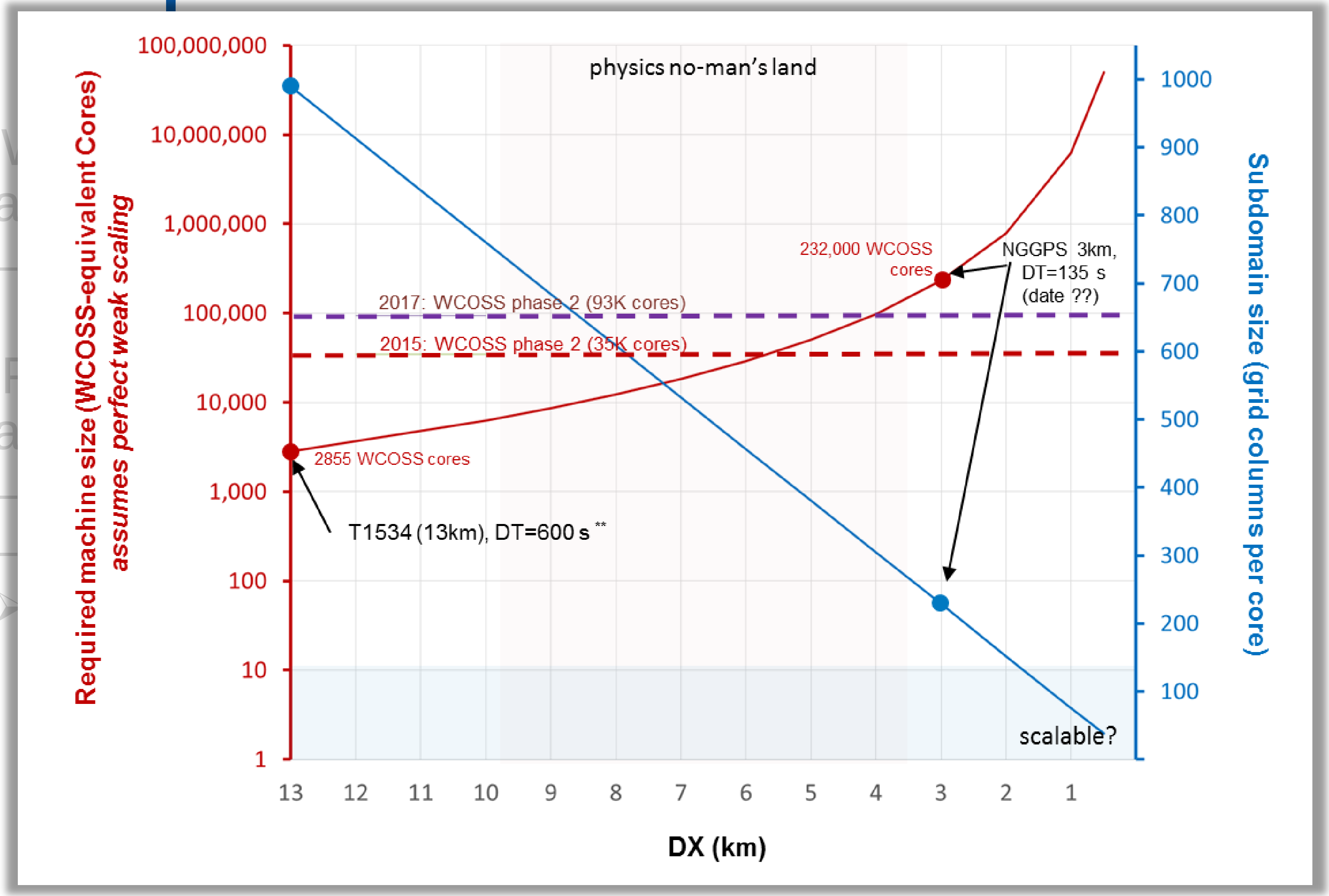
- NWP is one of the first HPC applications and, with climate, has been one its key drivers
  - Exponential growth in HPC capability has translated directly to better forecasts and steadily improving value to the public
- HPC growth continues toward Peta-/Exaflop, but only parallelism is increasing
  - More floating point capability
  - Proportionately less cache, memory and I/O bandwidth
  - ➢ **Parallelism scales in one fewer dimension than complexity**

HPC Imperative



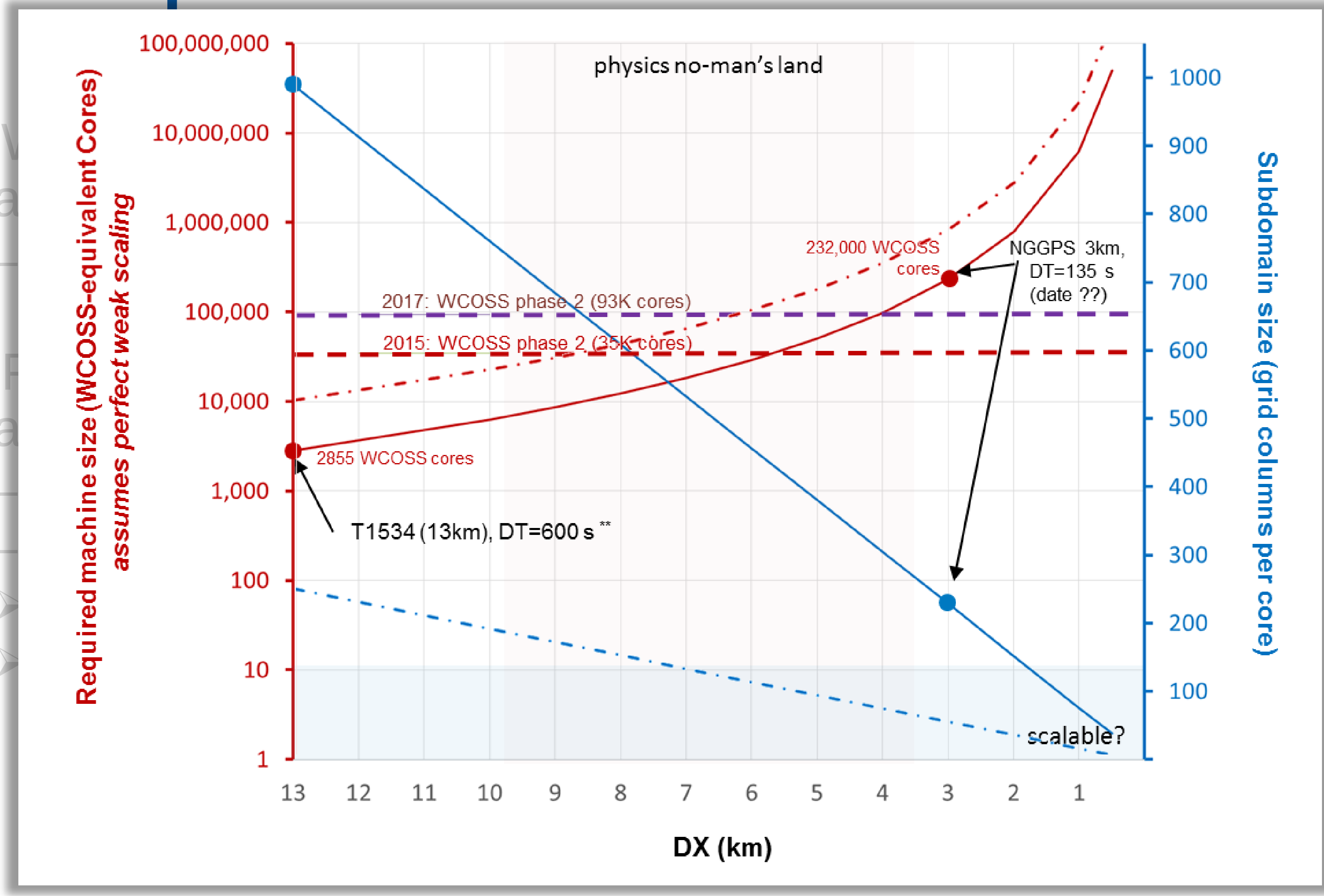
** T1534 GFS will be implemented operationally in Dec. 2014 to use 1856 WCOSS cores, 64 vertical levels and DT=450s. The plotted point is adjusted to 2855 cores, 128 levels and a 600 s time step to conform to planned higher-resolution configurations.

HPC Imperative



** T1534 GFS will be implemented operationally in Dec. 2014 to use 1856 WCOSS cores, 64 vertical levels and DT=450s. The plotted point is adjusted to 2855 cores, 128 levels and a 600 s time step to conform to planned higher-resolution configurations.

HPC Imperative

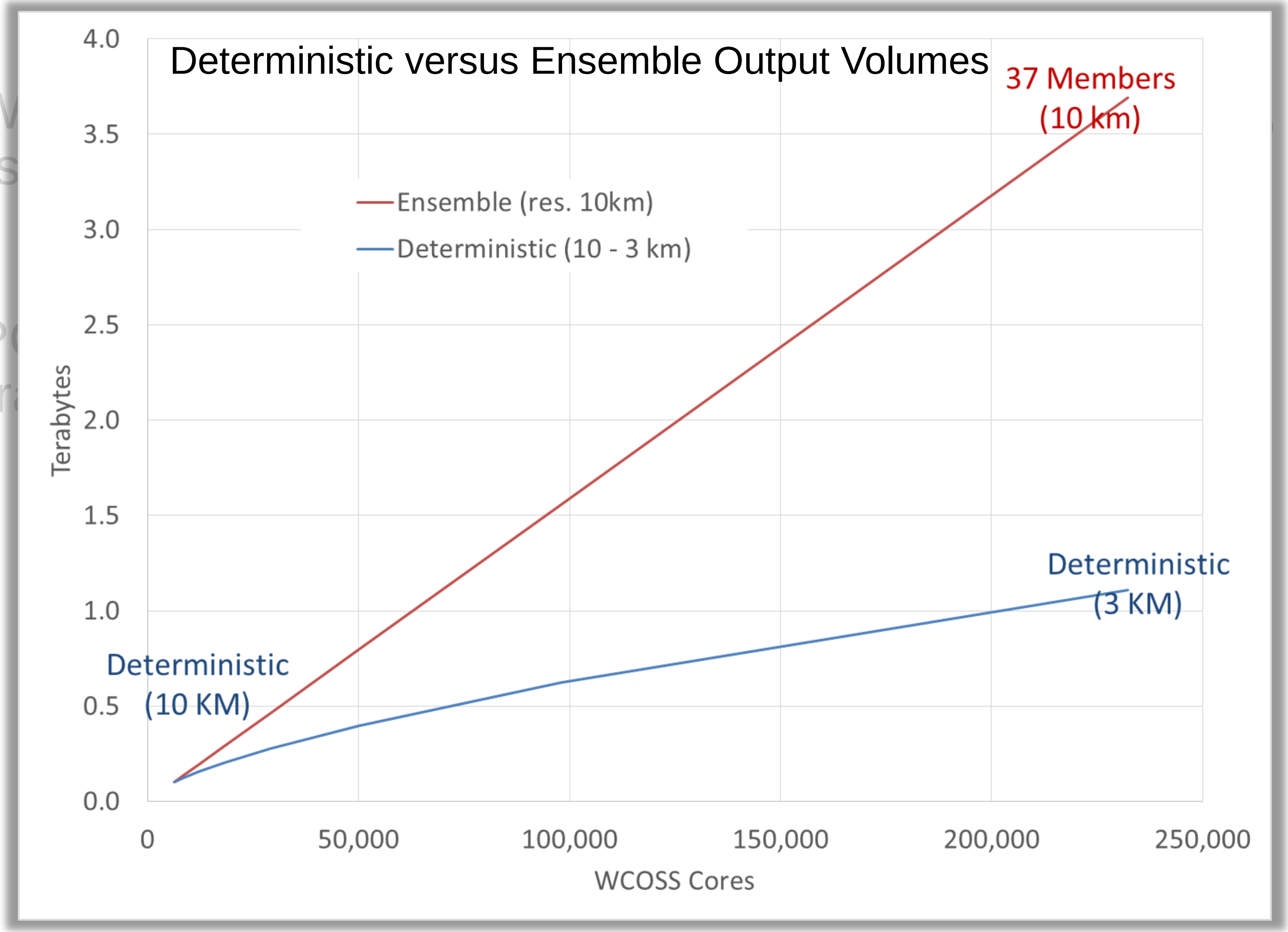


** T1534 GFS will be implemented operationally in Dec. 2014 to use 1856 WCOSS cores, 64 vertical levels and DT=450s. The plotted point is adjusted to 2855 cores, 128 levels and a 600 s time step to conform to planned higher-resolution configurations.

HPC Imperative

- NWP is one of the first HPC applications and, with climate, has been one its key drivers
  - Exponential growth in HPC capability has translated directly to better forecasts and steadily improving value to the public
- HPC growth continues toward Peta-/Exaflop, but only parallelism is increasing
  - More floating point capability
  - Proportionately less cache, memory and I/O bandwidth
  - Parallelism scales in one fewer dimension than complexity
  - ➢ **Ensembles scale computationally but move problem to I/O**

HPC Imperative



HPC Imperative

- NWP is one of the first HPC applications and, with climate, has been one its key drivers
  - Exponential growth in HPC capability has translated directly to better forecasts and steadily improving value to the public
- HPC growth continues toward Peta-/Exaflop, but only parallelism is increasing
  - More floating point capability
  - Proportionately less cache, memory and I/O bandwidth
  - NWP parallelism scales in one fewer dimension than complexity
  - Ensembles scale computationally but move problem to I/O
- Can operational NWP stay on the HPC train?
  - ➢ **Expose + exploit all available parallelism, especially fine-grain**
  - ➢ More scalable formulations

Next-Generation Global Prediction System

- Response to Hurricane Sandy (2012):
  - $14.8M / 5 year Research to Operations (R2O) Initiative
  - Produce state-of-the-art prediction system, **NGGPS**, by 2018
- Goals: Meet evolving national requirements over next 15-20 years
  - Global high-resolution weather prediction (3-10km)
  - High-impact weather: hurricanes, severe storms (0.5-2km nesting)
  - Extend skill to 30 days, seasonal climate
  - Coupled ocean-atmosphere-ice-wave modeling system
  - Ensemble forecasting and data assimilation
  - Aerosol forecasting, others
- Needed
  - ➢ *New non-hydrostatic dynamics scalable to O(100K) cores*

Next-Generation Global Prediction System

Task: New Scalable Non-hydrostatic Dynamical Core

(*in 5 years*??)

- Use a model already under development
  - Coordinate with HIWPP program (Tim Schneider's talk)
  - Select from 5 candidate models + current system:

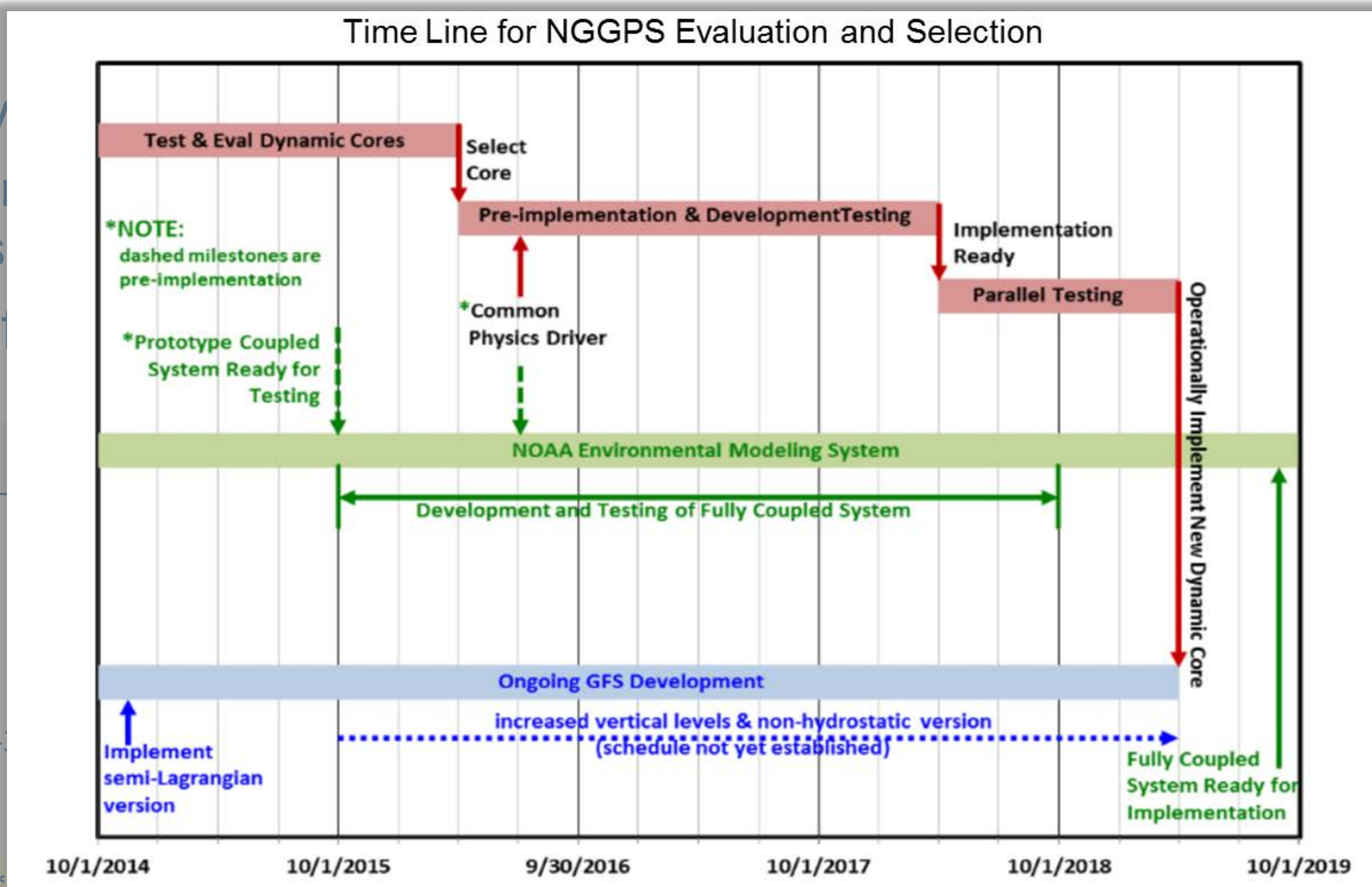
Model	Organization	Numeric Method	Grid
NIM	NOAA/ESRL	Finite Volume	Icosahedral
MPAS	NCAR/LANL	Finite Volume	Icosahedral/Unstructured
NEPTUNE	Navy/NRL	Spectral Element	Cubed-Sphere with AMR
HIRAM/FV-3	NOAA/GFDL	Finite Volume	Cubed-Sphere, nested
NMMB	NOAA/EMC	Finite difference/Polar Filters	Cartesian, Lat-Lon
GFS-NH **	NOAA/EMC	Semi-Lagrangian/Spectral	Reduced Cartesian

** current operational baseline, non-hydrostatic option under development

Next-Generation Global Prediction System

- Non-hy
  - Coo
    pres
- Select

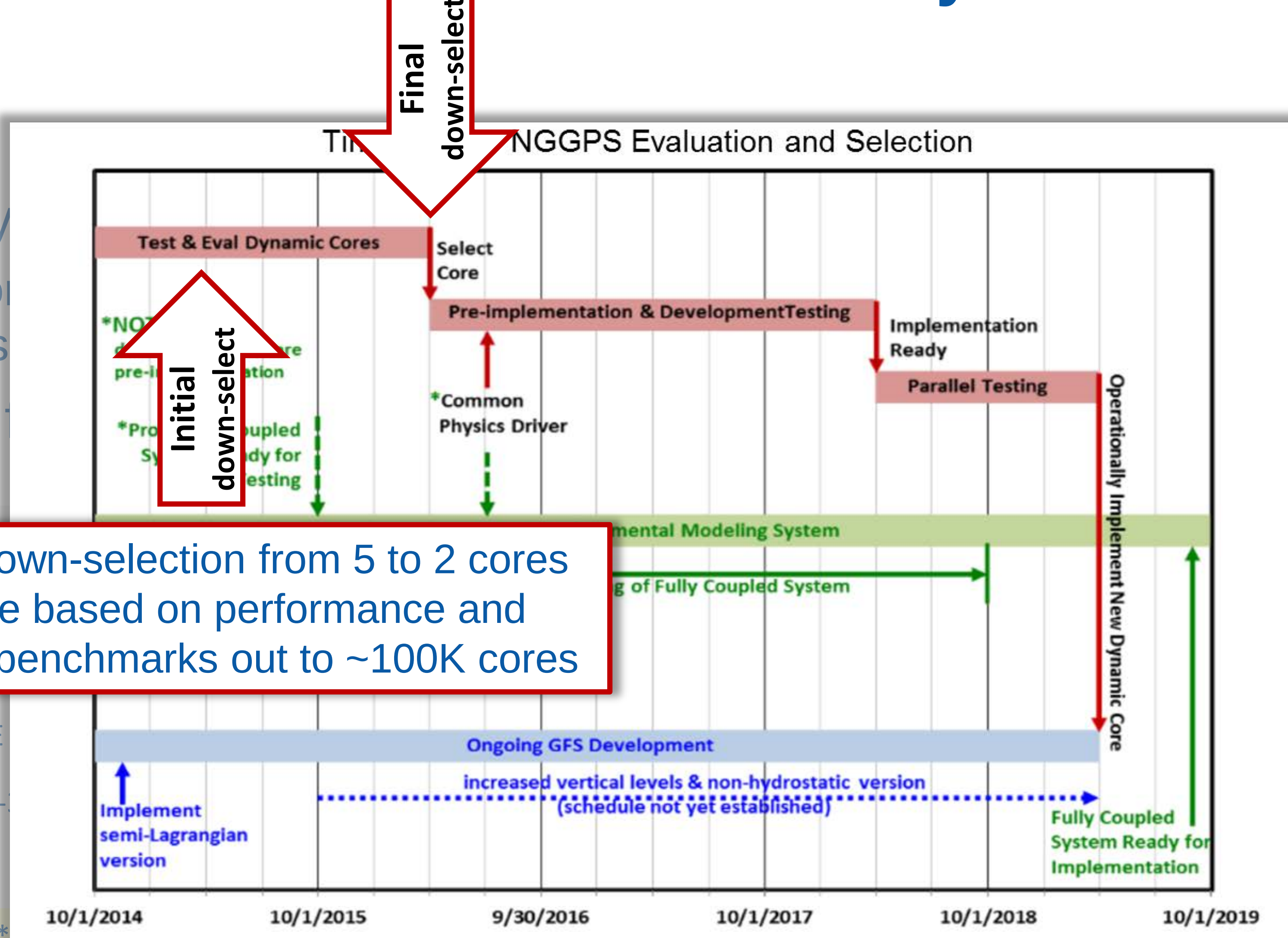
Model
NIM
MPAS
NEPTUNE
HIRAM/FV-3
NMMB
GFS-NH **



** current operational baseline, non-hydrostatic option under development

Next-Generation Global Prediction System

- Non-hy
  - Coo
    pres
- Select



Initial down-selection from 5 to 2 cores will be based on performance and scaling benchmarks out to ~100K cores

NEPTUNE

HIRAM/FV-

NMMB

GFS-NH **

** current operational baseline, non-hydrostatic option under development

NGGPS Level-1 Benchmarking Plan

- Advanced Computing Evaluation Committee:
  - Co-Chairing with Mark Govett (NOAA/ESRL)
- Investigate and report near-term and lifetime prospects for performance and scaling of NGGPS candidate models
  - Test case:
    - Idealized global baroclinic wave
    - Monotonically constrained advection of ten tracer fields
    - Artificially generated initial “checkerboard” patterns
  - Two workloads:
    - 13km “performance” workload – resources needed to meet operational requirements (8.5 minutes/day)
    - 3km workload measure scalability out to ~150K cores
  - All conventional multi-core (no accelerators)
- Verification
  - Reproduce baseline solution statistics for each model
  - Return output from runs for additional verification and validation
- Additional evaluation of software design and readiness for HPC

NGGPS Level-1 Benchmarking Plan

Benchmark systems

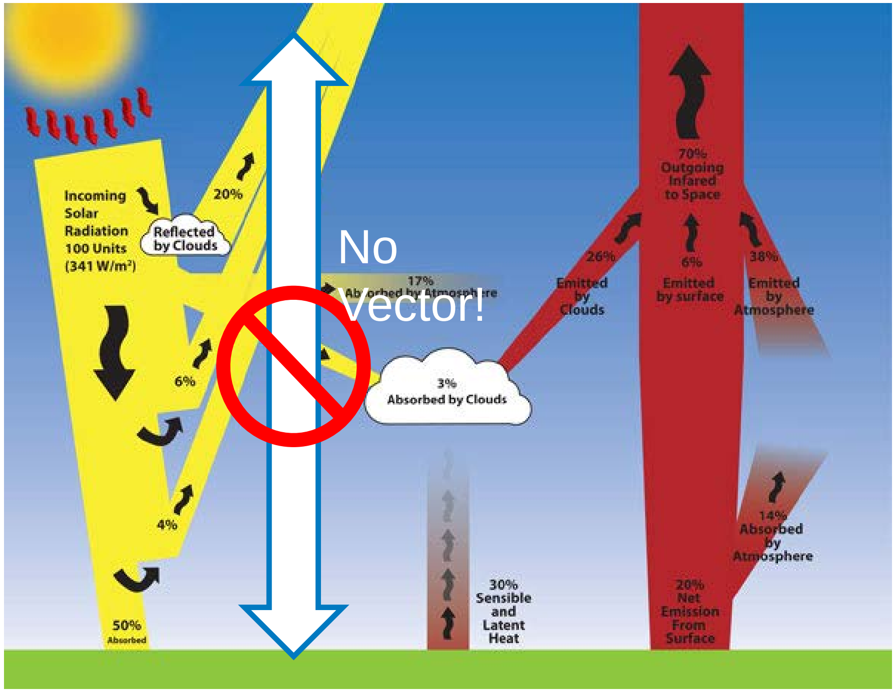
- **Edison:** National Energy Research Scientific Computing Center (DOE/BNL)
  133,824 cores, Xeon Ivy Bridge, 24 cores per node
  Four million discretionary hours awarded
- **Stampede**: Texas Advanced Computing Center (NSF)
  102,400 cores, Xeon Sandy Bridge), 16 cores per node
- **Pleiades**: NASA/Ames Research Center
  108,000 cores, Xeon Ivy Bridge, 20 cores per node
  Possibility of ~100,000 cores of Xeon Haswell by benchmarking time
- **Yellowstone**: National Center for Atmospheric Research (NSF)
  72,000 cores, Xeon Sandy Bridge, 16 cores per node

Scaling Operational NWP (Summary)

- Will have NGGPS Level-1 Benchmark results Spring ‘15
- But we know the long-term future for deterministic global forecast models:
  - Scaling will *eventually* run out
  - We aren’t there yet
- Which models make the most of headroom there is:
  - Chose models with the best weak scaling characteristics
    - Don’t give up anything on number of time-steps per second
    - Nearby communication patterns instead of non-local
  - Take longest time step possible
    - Semi-Lagrangain gives 5x DT, but trades accuracy for stability
    - Do skill improvements translate to convective scales?
    - Need for embedded high-res. models: nesting or Panta Rhei approach
  - Make effective core speeds faster
    - Exploit more parallelism: vertical, tracers, **and esp. *fine-grained***

Effect of Fine-grained optimization on RRTMG* radiative transfer physics

- Accurate calculation of fluxes and cooling rates from incoming (shortwave) and outgoing (longwave) radiation
- Used in many weather and climate models
  - NCAR WRF and MPAS
  - NCAR CAM5 and CESM1
  - NASA GEOS-5
  - NOAA NCEP GFS, CFS, RUC
  - ECMWF
- Significant computational cost
  - Coded as 1-D vertical columns but poor vectorization in this dimension



One column of a weather or climate model domain

https://www.aer.com/science-research/atmosphere/radiative-transfer

(*Iacono et al. JGR, 2008; Mlawer et al., JGR, 1997)

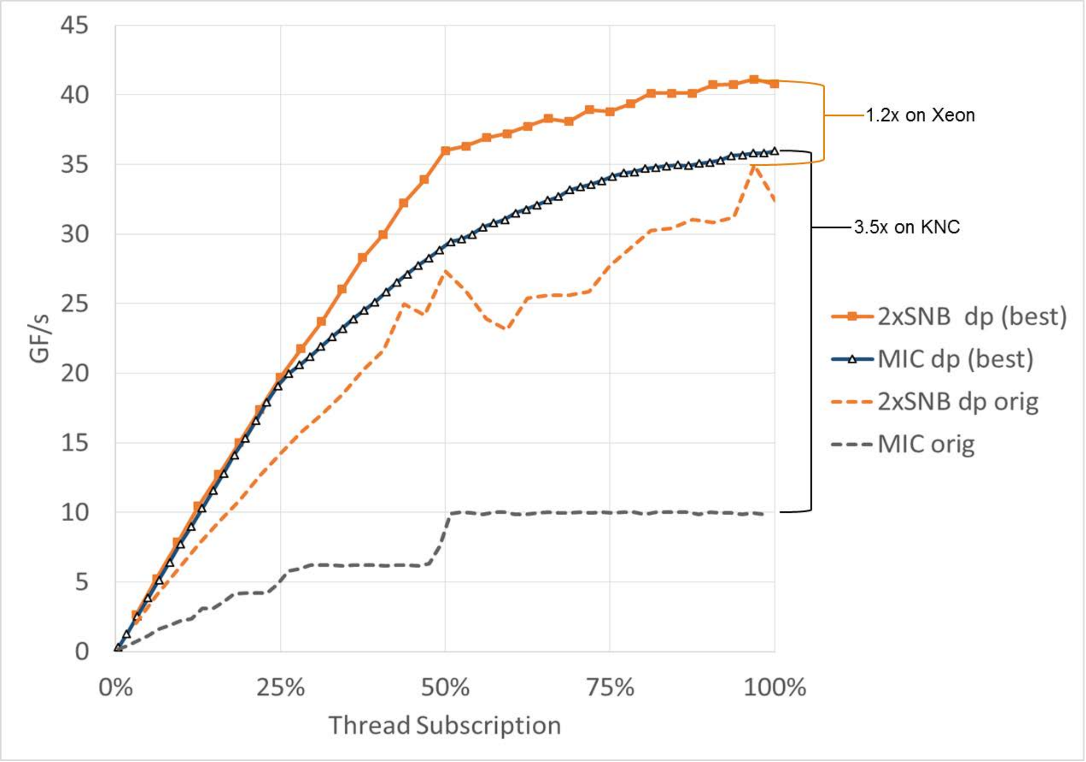
Performance results: RRTMG Kernel on Xeon Phi and host Xeon (SNB)

Workload

- 1 node of 80 node NMMB run
  - 4km CONUS domain
- 1 RRTMG invocation
  - 18819 columns, 60 levels
  - 46.5 billion DP floating point ops

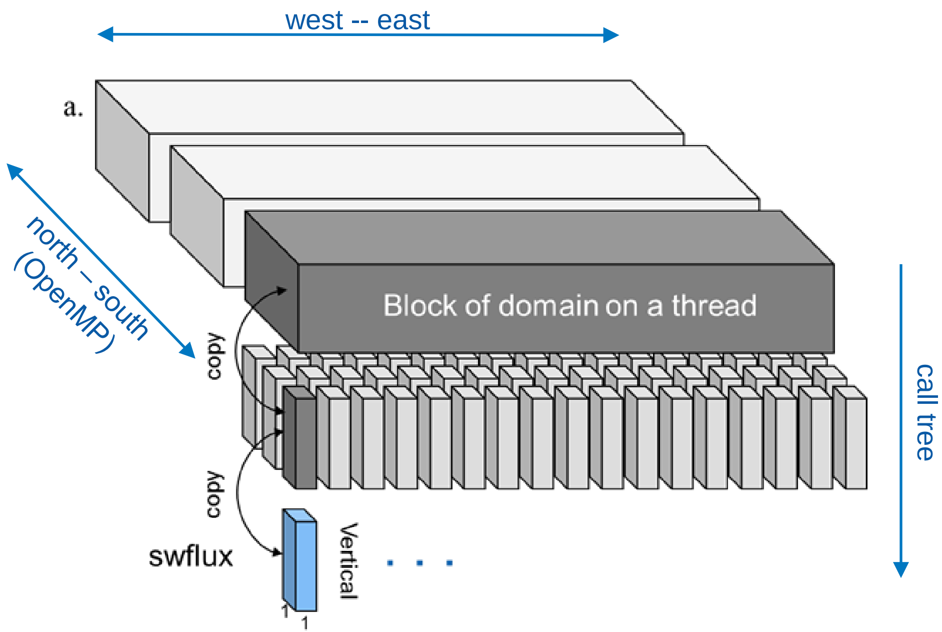
Code restructuring

- Increase concurrency
- Increase vectorization
- Decrease memory system pressure
- Performance improves on host too



Restructuring RRTMG in NMM-B

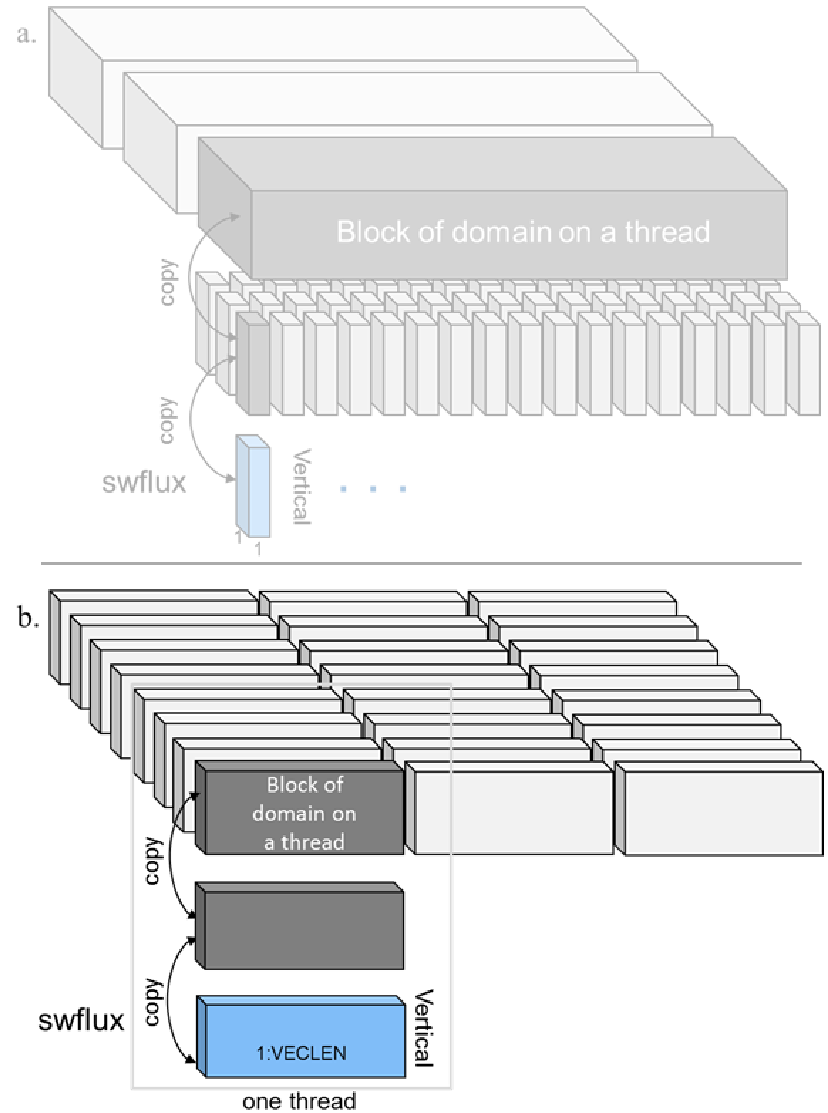
- Concurrency and locality
  - Original RRTMG called in OpenMP threaded loop over South-North dimension



- Vectorization
  - Originally vertical pencils

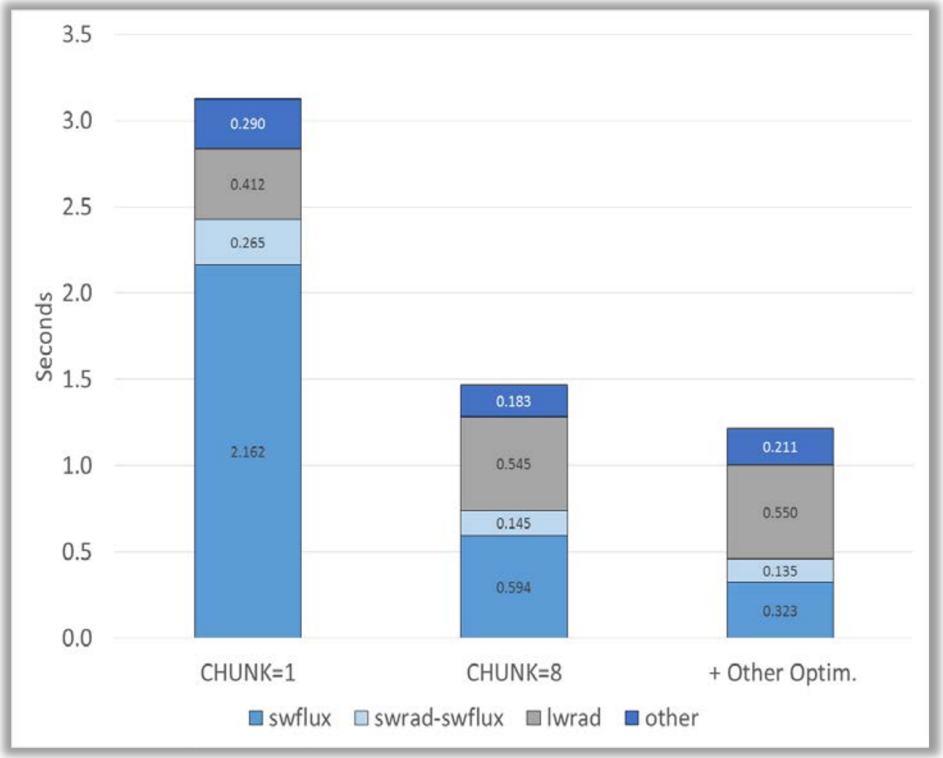
Restructuring RRTMG in NMM-B

- Concurrency and locality
  - Original RRTMG called in OpenMP threaded loop over South-North dimension
  - Rewrite loop to iterate over tiles in two dimensions
  - Dynamic thread scheduling
- Vectorization
  - Originally vertical pencils
  - Extend inner dimension of lowest-level tiles to width of SIMD unit on KNC
  - Static definition of VECLEN



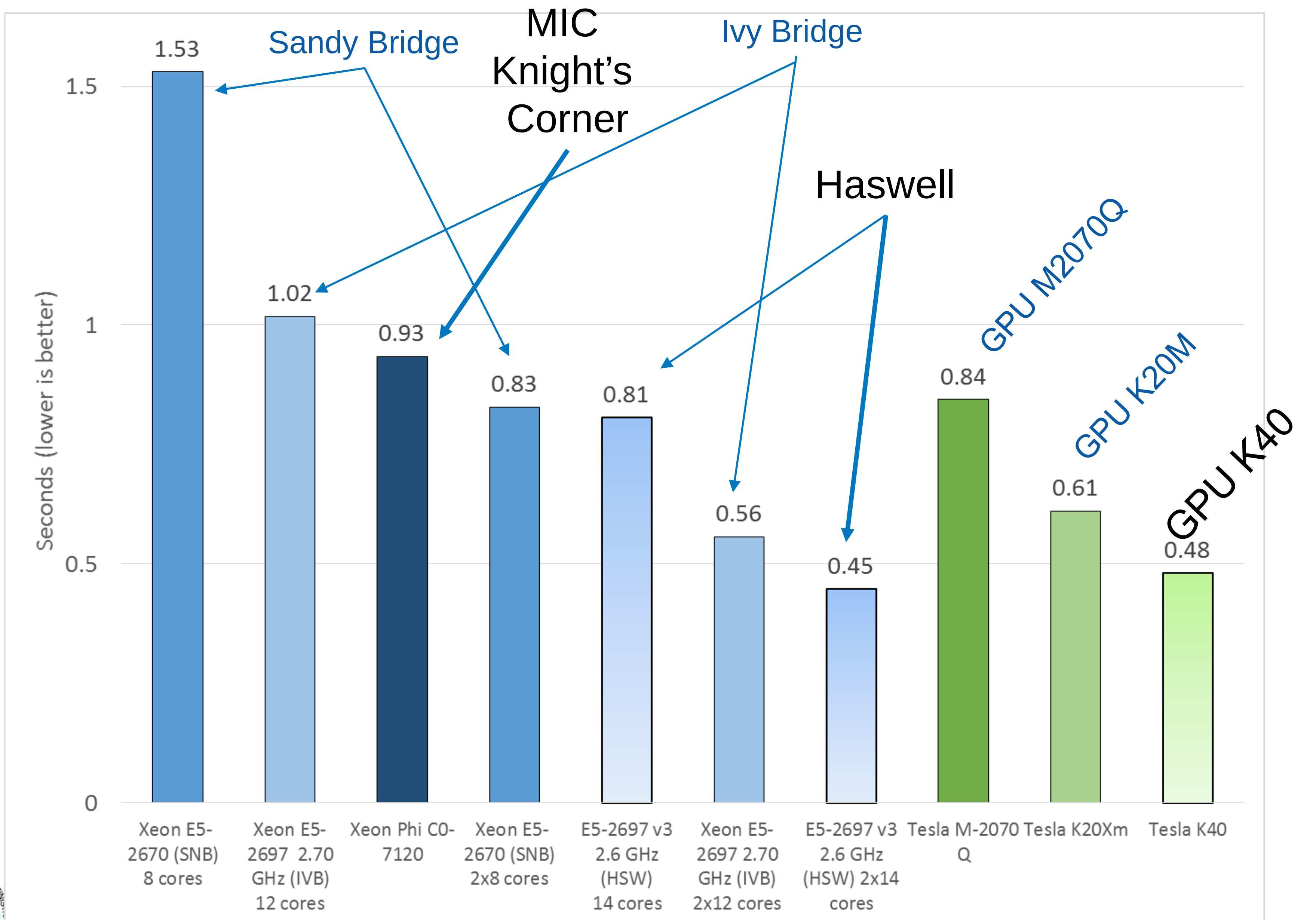
Effect of Fine-grained optimization on RRTMG radiative transfer physics

- Improvement
  - 2.8x Overall
    - 5.3x in SWRAD
    - 0.75x in LWRAD *(degraded)*
- Increasing chunk size results in
  - 2.5x increase in working set size from 407KB to 1034KB per thread
  - 4x increase in L2 misses
- Memory traffic
  - Increased from 59 to 124 GB/s, still short of saturation
  - Key bottlenecks
    - Memory latency
    - Instruction pipeline stalls around hardware division instruction



Michalakes, Iacono, Jessup. Optimizing Weather Model Radiative Transfer Physics for Intel's Many Integrated Core (MIC), Architecture. Preprint.
http://www.Michalakes.us/michalakes_2014_web_preprint.pdf

Comparison to GPU Performance (32-bit; shortwave)



Outlook for accelerators

- For now, neither GPU nor current MIC generation are compelling compared with conventional multi-core Xeon
  - Improving performance on MIC leads to faster Xeon performance
- Next release of Xeon Phi: Knights Landing
  - Hostless, no PCI gulf
    - NERSC's "Cori" system (mid 2016): 9,300 single socket KNL nodes
  - On-package memory
    - 5x Stream Triad bandwidth over DDR4
  - More powerful cores
    - Out-of-order, advanced branch prediction, AVX-512 ISA
    - Overall 3x faster single-thread performance (workload dependent)
  - Other improvements (NDA)

Outlook for NWP on HPC

- Deterministic forecasting will stall eventually but still has headroom
- Recast or develop new modeling systems that emphasize parallelism and locality
- Continue to investigate hardware and programming models that provide highest possible flops per second-dollar-watt
- Increased computing power will continue to add value through other approaches (ensembles, data assimilation, coupled systems)