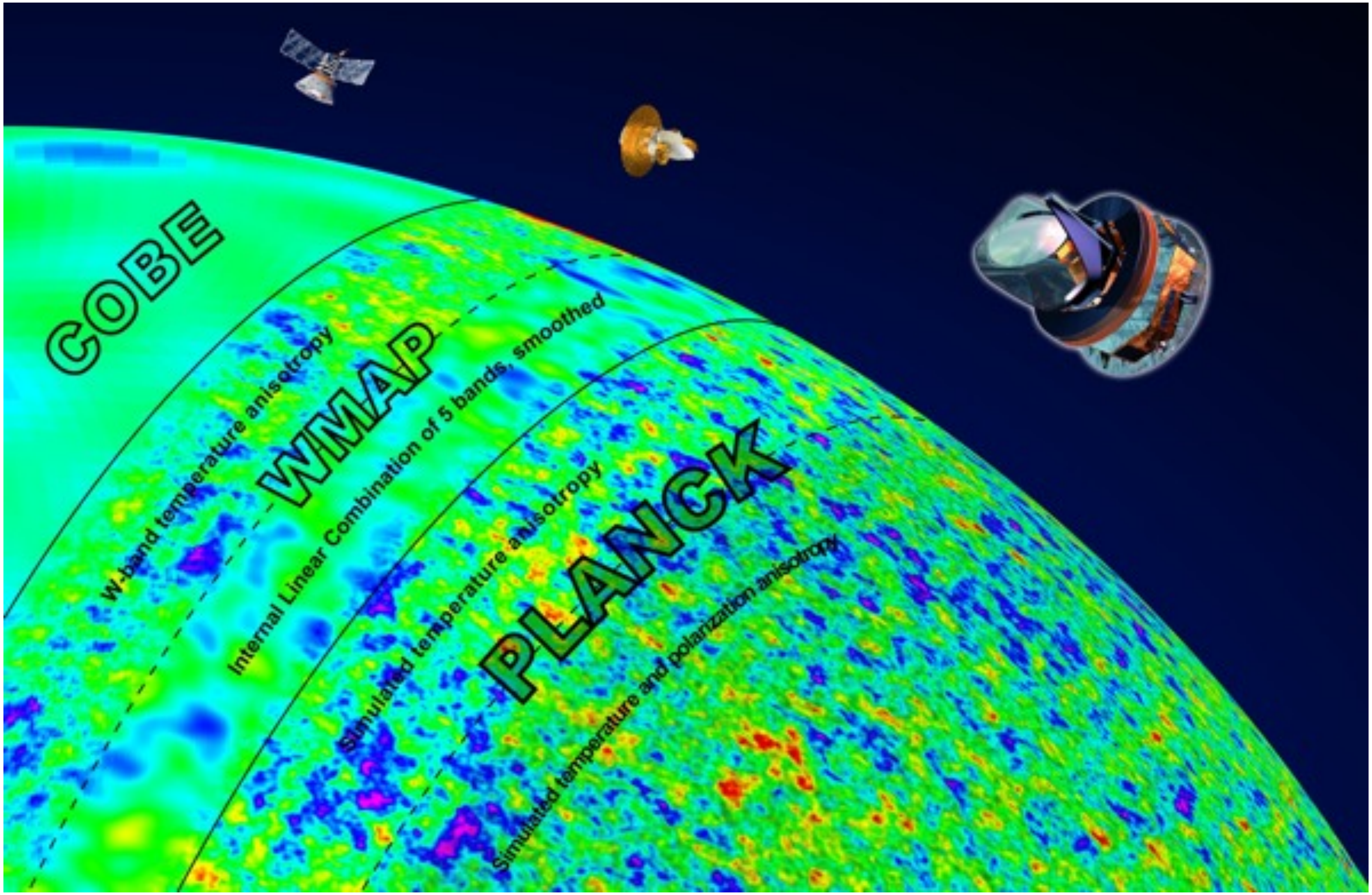
Cosmic Microwave Background Data Analysis At The Petascale And Beyond

Julian Borrill
Computational Cosmology Center, LBNL
& Space Sciences Laboratory, UC Berkeley

with Chris Cantalupo, Ted Kisner, Radek Stompor & Rajesh Sudarsan
and the BOOMERanG, EBEX, MAXIMA, MAXIPOL, Planck &

Evolution of CMB Observations



Ongoing quest for higher resolution & intrinsically fainter (polarization) signals.

Evolution of CMB Data Sets

Increased resolution & sensitivity required for evolving science goals
implies ever larger data sets to achieve the necessary S/N.

Date	Experiment	Description	Time Samples (N_t)	Sky Pixels (N_p)
1990-93	COBE	All-sky, low-res, T	8×10^{8}	3×10^{3}
1998	BOOMERaNG	Cut-sky, mid-res, T	9×10^{8}	3×10^{5}
2001-10	WMAP	All-sky, mid-res, TE	2×10^{11}	6×10^{6}
2009-11	Planck	All-sky, high-res, TE	5×10^{11}	1×10^{8}
2011-13	PolarBeaR	Cut-sky, high-res, TEB	3×10^{13}	1×10^{7}
2012-15	QUIET-II	Cut-sky, high-res, TEB	1×10^{14}	7×10^{5}
~ 2020-25	CMBpol (EPIC,CoRE)	All-sky, high-res, TEB	4×10^{14}	6×10^{8}

Number of samples increases 1000-fold over the next 15 years –

Simulation & Map-Making

- The most computationally expensive step in CMB data analysis is generating Monte Carlo realizations of the data.
  - For each realization:
    - Simulate the time-ordered data.
    - Solve for the map of that data.

- Challenges for algorithms & implementations:
  - Achievable scaling: (log-)linear algorithms.
  - Minimal prefactor: smart pre-conditioners etc.
  - Maximal efficiency: data delivery.

Optimization

- Targeting a particular CMB dataset implies strong scaling.
- IO & communication performance don't grow like calculation.
- Exacerbated by (log)-linear algorithms performing few flop/byte.
- Trade-offs between sub-systems
  - Replace IO with communication & calculation.
  - Replace communication with calculation.
  - Exact trade-offs may be system, concurrency and problem dependent.
  - For strong scaling this has to be repeated at each generation.
- Can lead to ambiguity in sub-system performance

MADmap

- MADmap solves for the maximum likelihood pixelized sky map (m_p) given the time-ordered data (d_t) & pointing solution (A_{tp}) and piecewise stationary Gaussian noise correlations ($N_{tt'}$):

$$m = (A^T N^{-1} A)^{-1} A^T N^{-1} d$$

- MAXIMA/BOOMERanG: use exact solution with explicit pixel-domain matrix operations, scaling as $O(N_p^3)$
  - High efficiency & low prefactor (Level3 BLAS).
  - Crippling cubic scaling.

- Planck onwards: approximate iterative solution with implicit time-domain matrix operations, scaling as $O(N_{it} N_t)$
  - Lower efficiency & higher prefactor (FFT+PCG).
  - Manageable log-linear scaling.

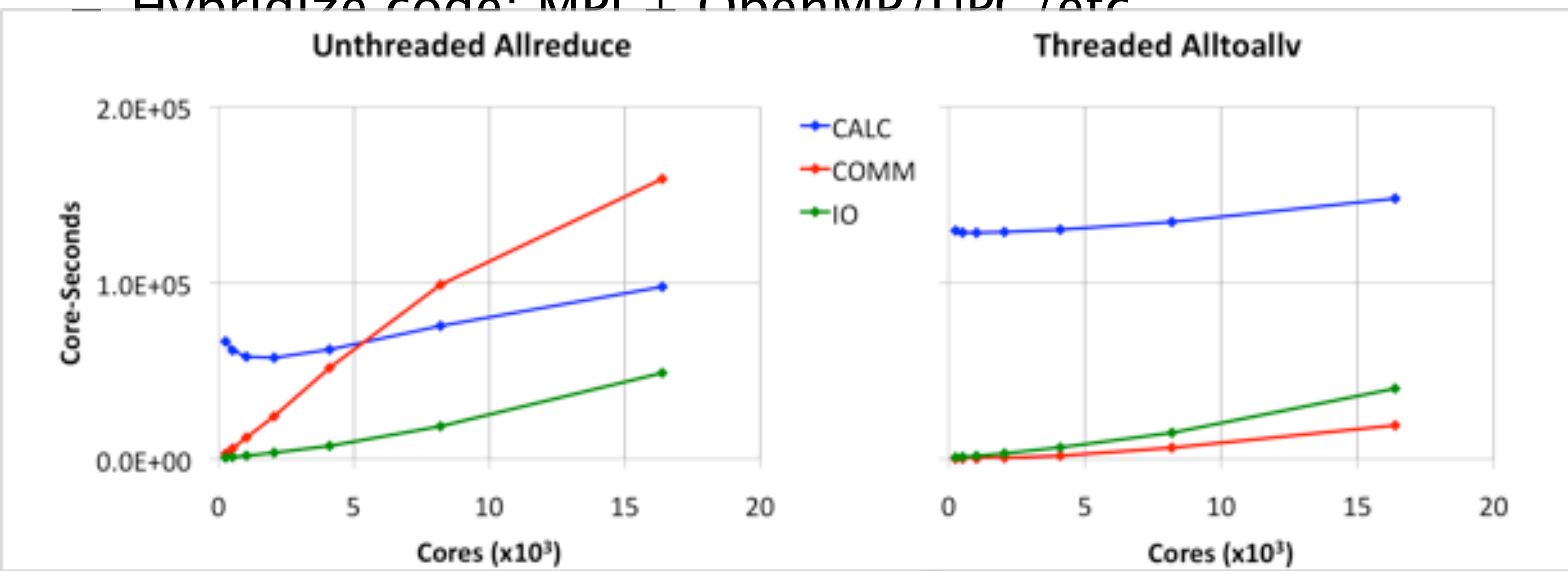
The I/O Bottleneck

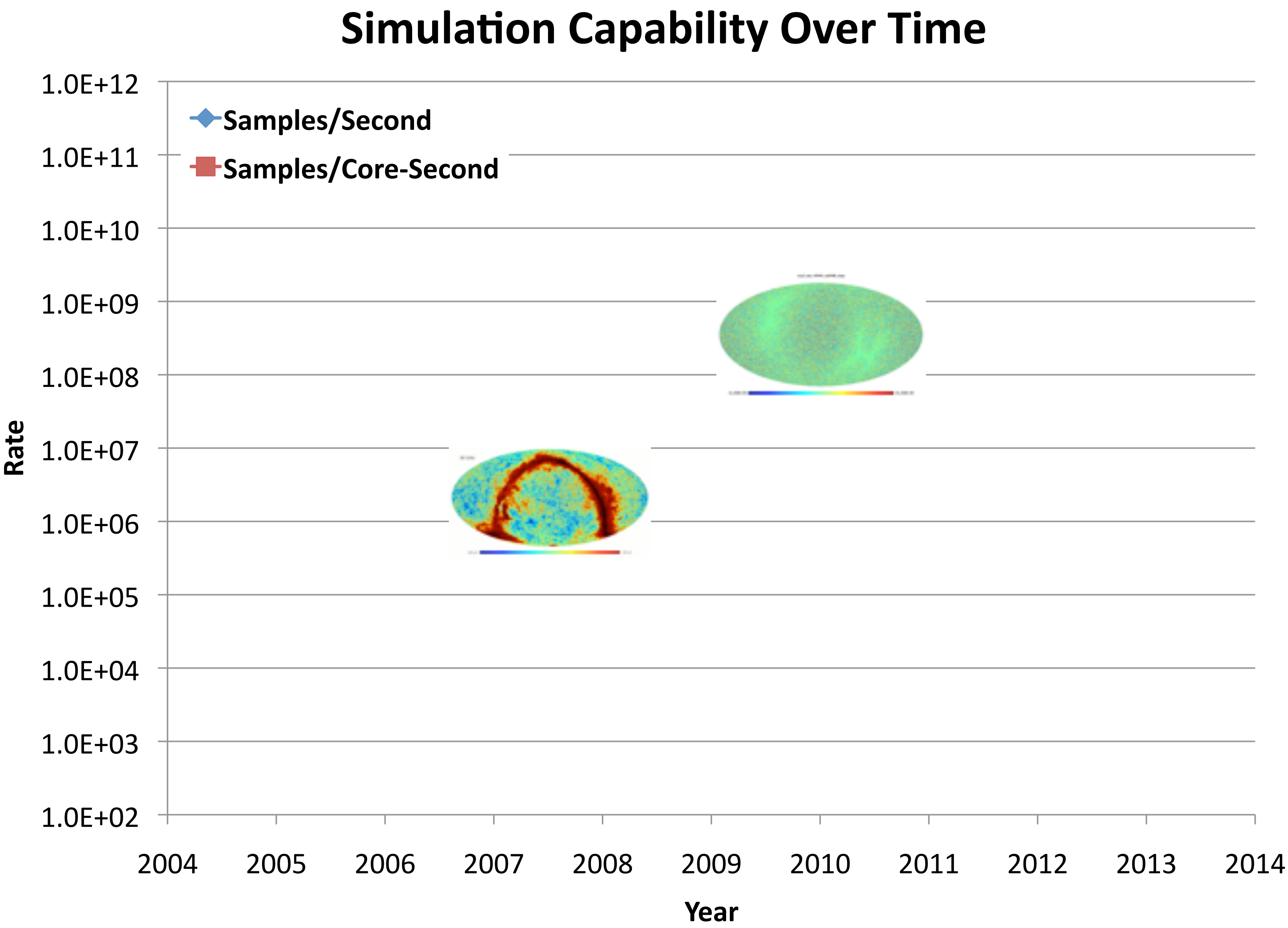
- Traditional approach:
  - Simulation:
    - Read pointing for each detector: 3 N_t
    - Write timestream for each detector: N_t
  - Mapping:
    - Read pointing & timestream for each detector: 4 N_t
    - Write map: N_p

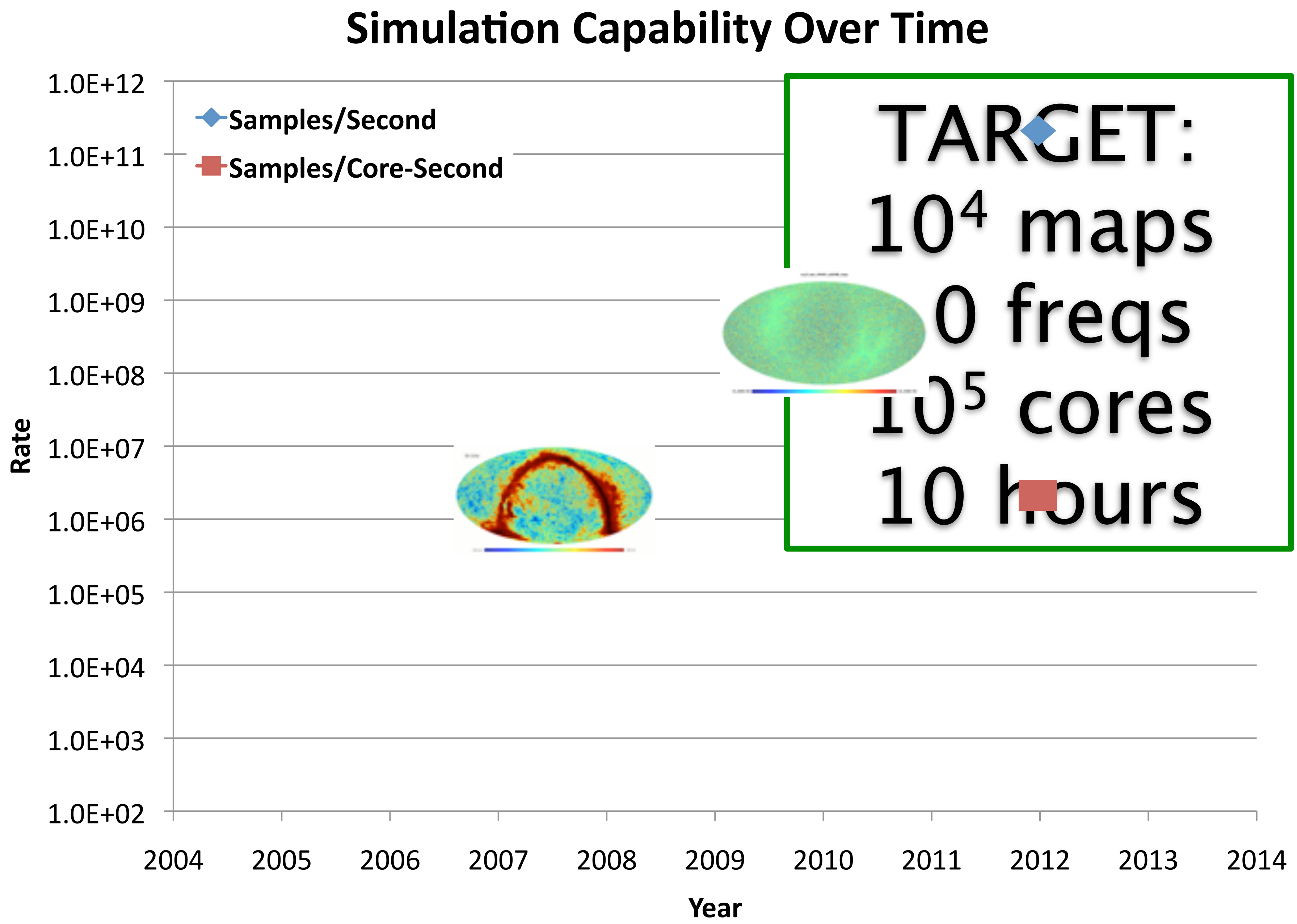
- Optimizations:
  - On-the-fly pointing reconstruction
    - Calculate dense detector from sparse boresight pointing.
  - On-the-fly simulation
    - Generate simulation on demand during map-making.
  - Reduces both read and write volumes by O(100 N_d)

The Communication Bottleneck

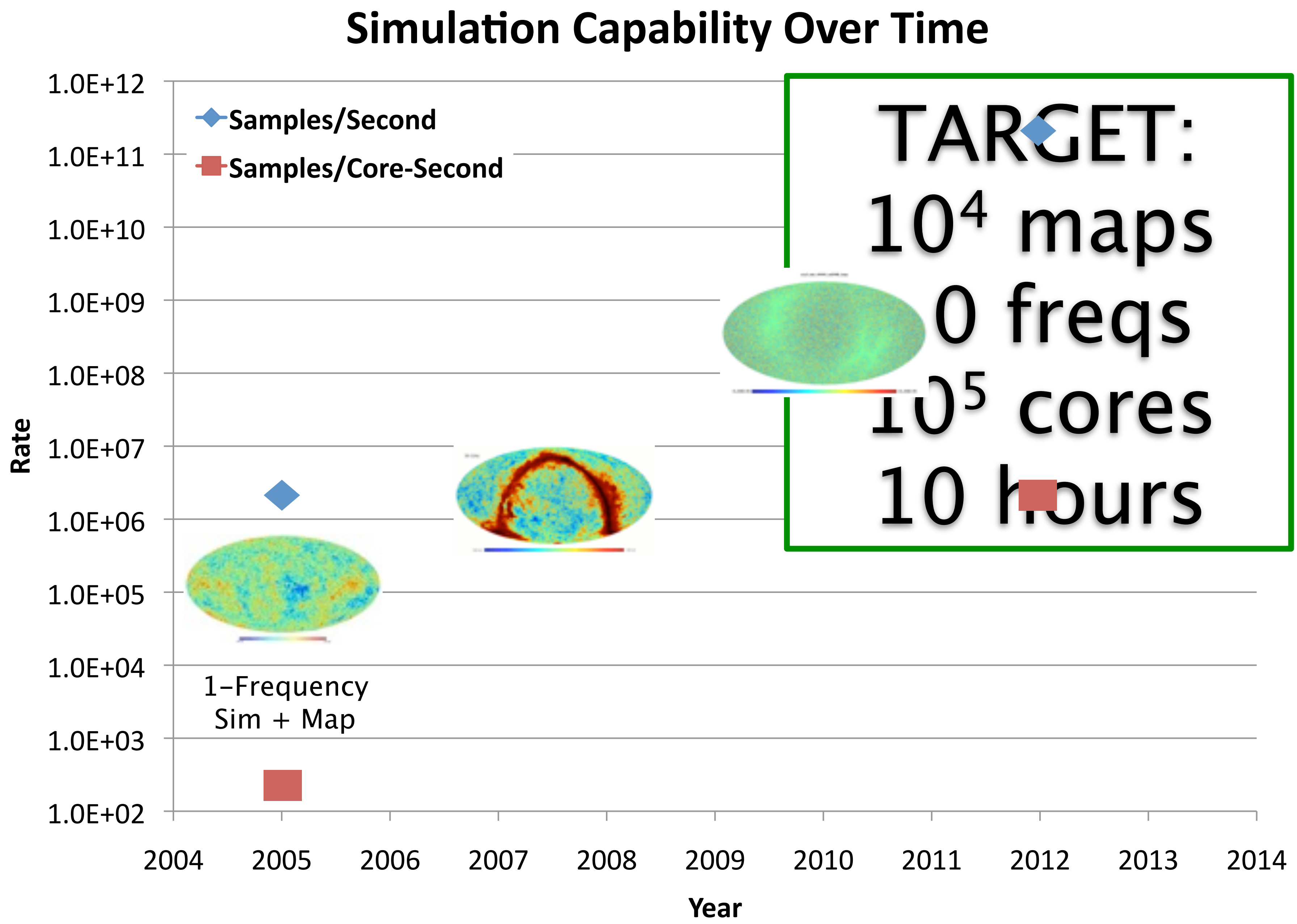
- To load-balance the calculation, time samples are distributed evenly across the processes; each reduces its samples to generate a sub-map.
- At each PCG iteration all sub-maps must be combined across all processes.
- Simplest approach uses MPI_Allreduce, but this does not scale well.
- Optimizations:
  – Hybridize code: MPI + OpenMP/UPC/etc.

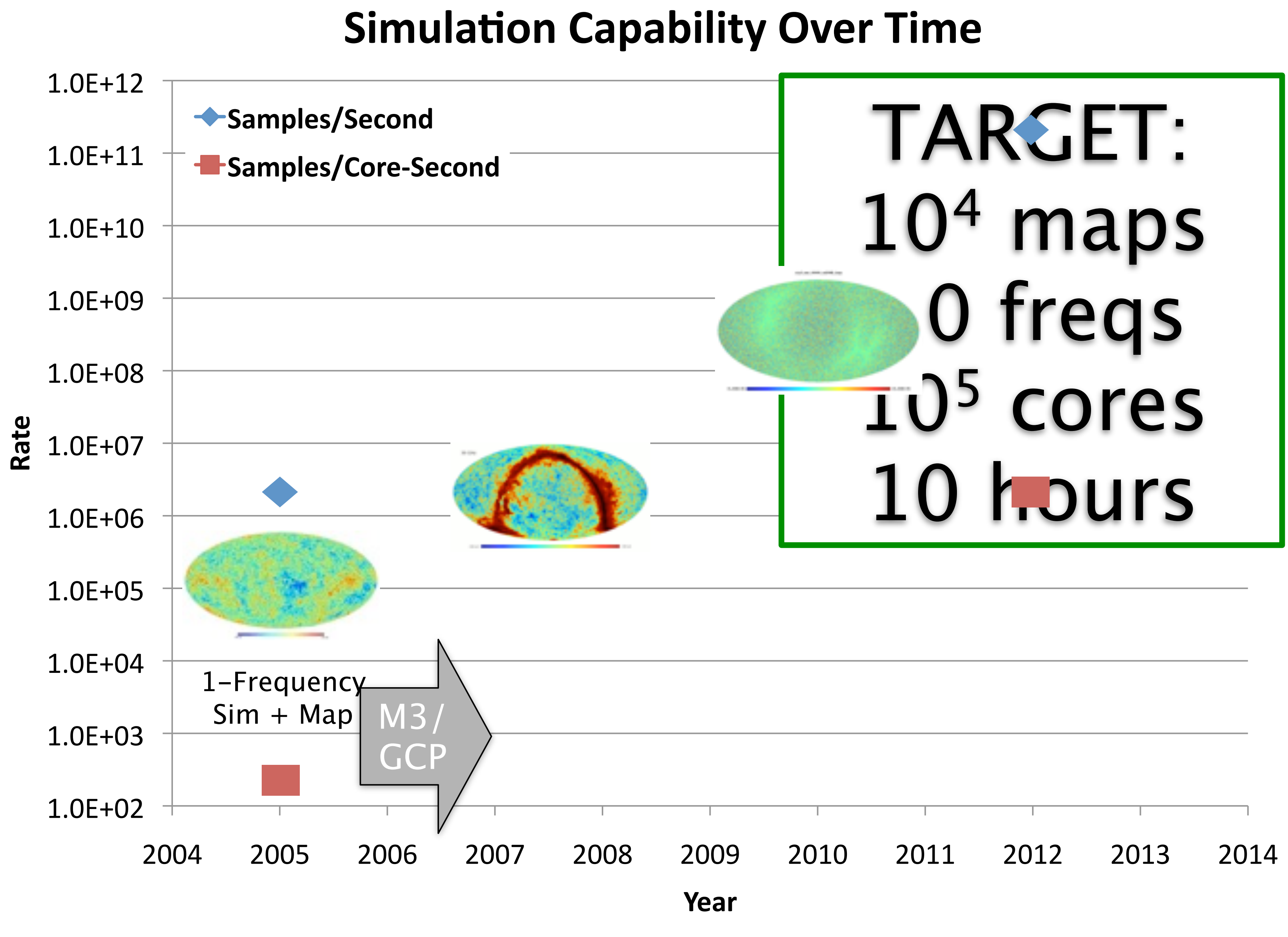


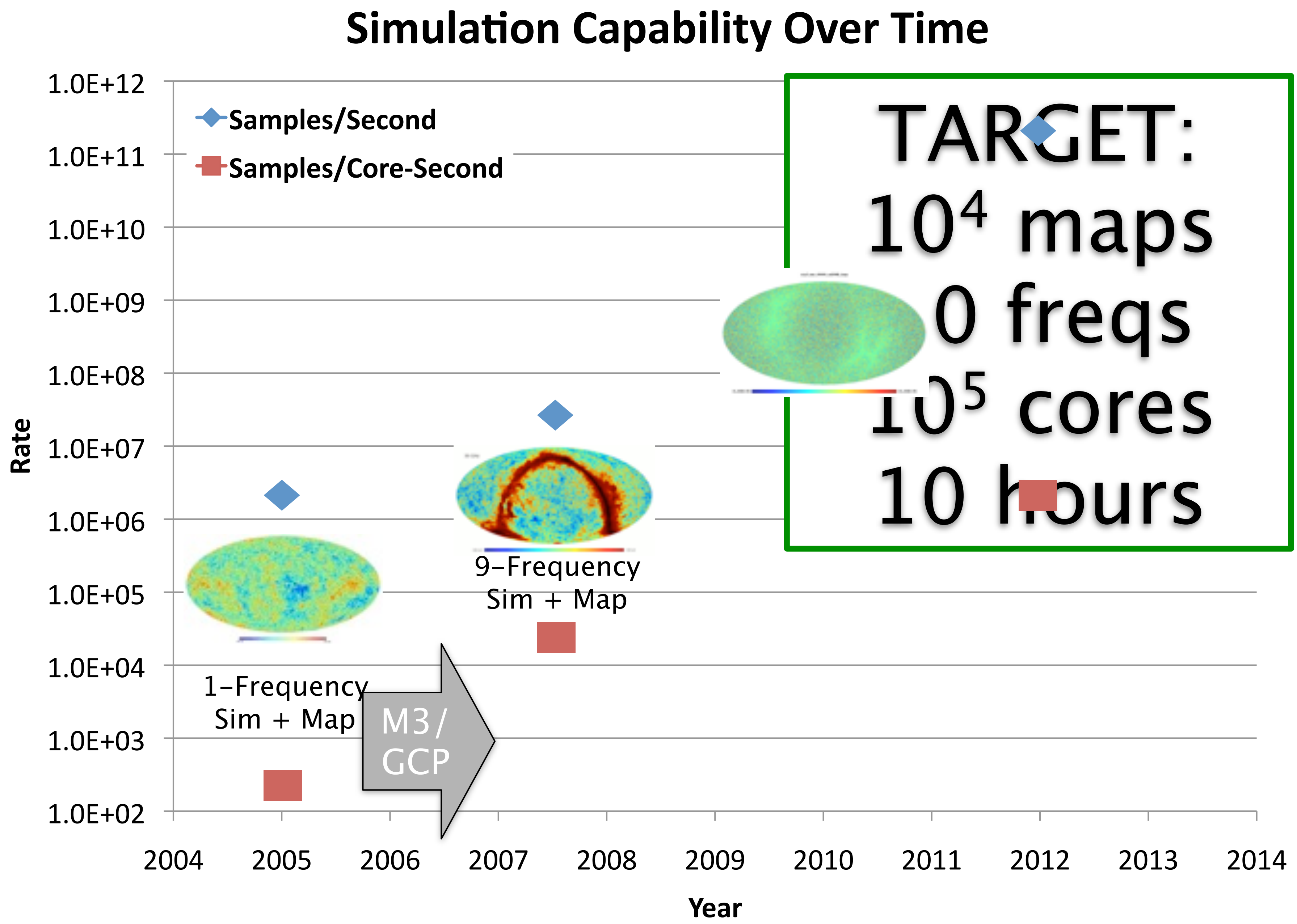

Simulation Capability Over Time
Samples/Second
Samples/Core-Second
1.0E+12
1.0E+11
1.0E+10
1.0E+09
1.0E+08
1.0E+07
1.0E+06
1.0E+05
1.0E+04
1.0E+03
1.0E+02
Rate
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
Year

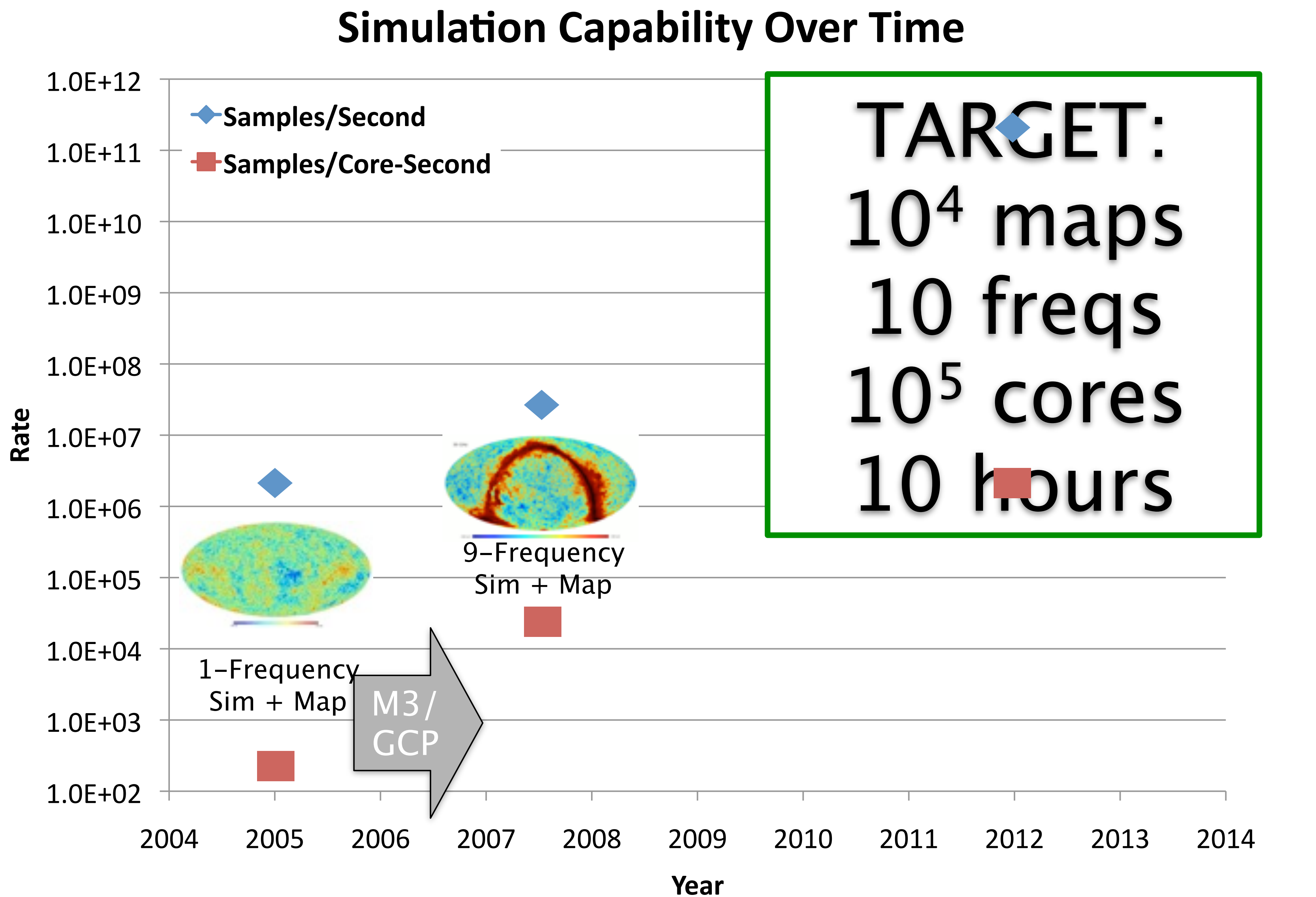


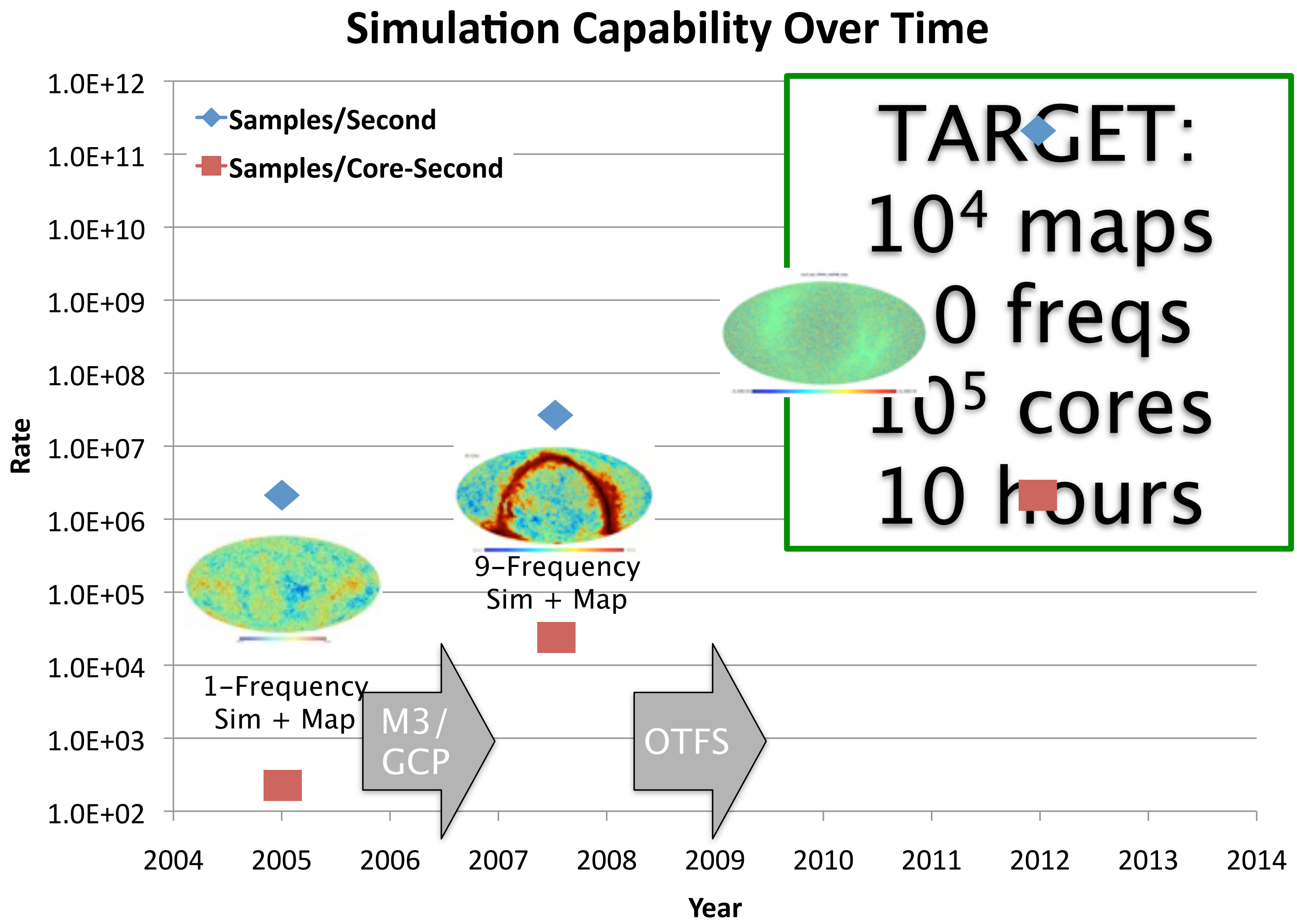
Simulation Capability Over Time
Samples/Second
Samples/Core-Second
1.0E+12
1.0E+11
1.0E+10
1.0E+09
1.0E+08
1.0E+07
1.0E+06
1.0E+05
1.0E+04
1.0E+03
1.0E+02
Rate
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
Year
TARGET:
10^4 maps
0 freqs
10^5 cores
10 hours


Simulation Capability Over Time
Samples/Second
Samples/Core-Second
TARGET:
10⁴ maps
0 freqs
10⁵ cores
10 hours
1-Frequency
Sim + Map
1.0E+12
1.0E+11
1.0E+10
1.0E+09
1.0E+08
1.0E+07
1.0E+06
1.0E+05
1.0E+04
1.0E+03
1.0E+02
Rate
2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
Year

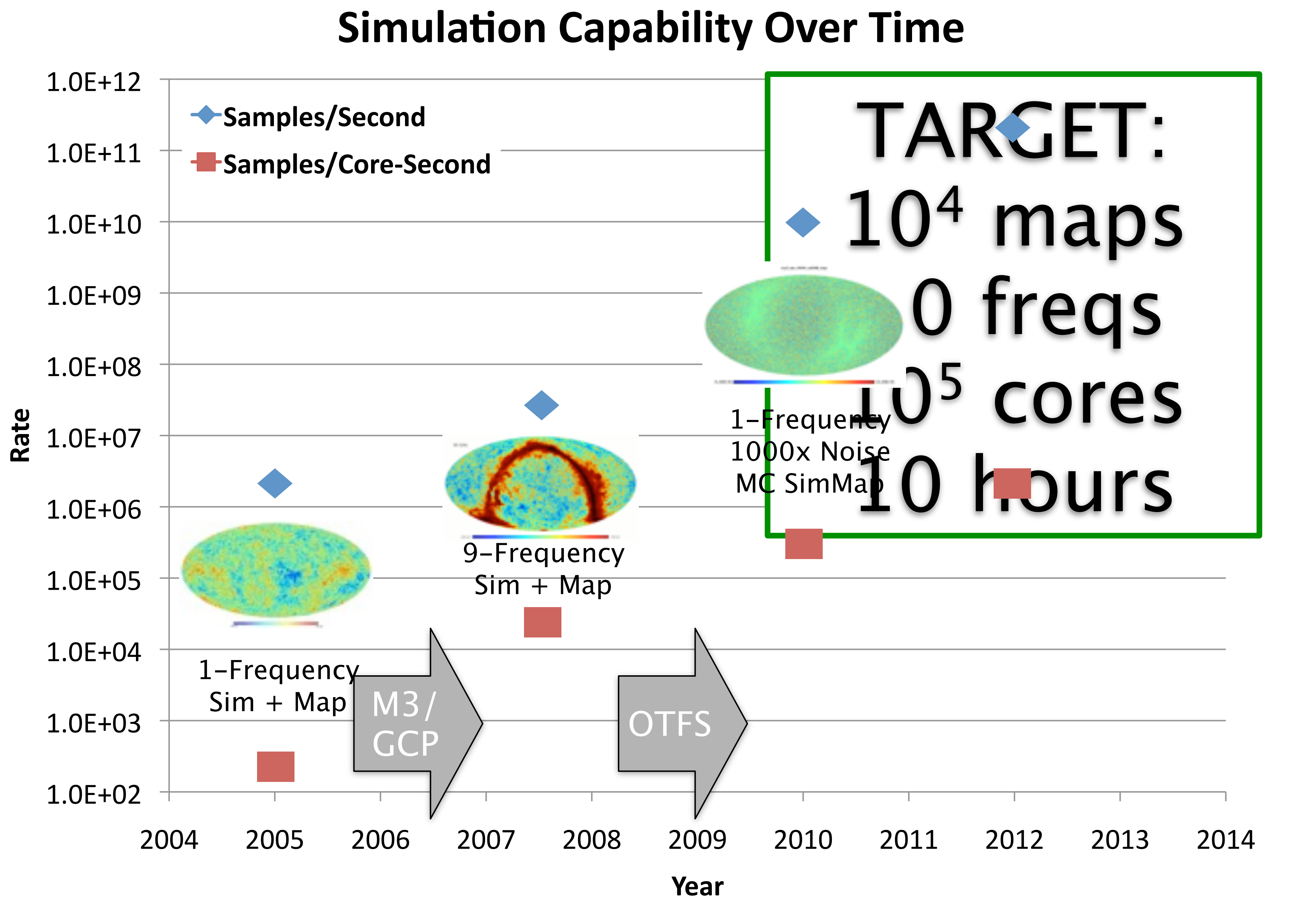

Simulation Capability Over Time
Samples/Second
Samples/Core-Second
TARGET:
10^4 maps
10 freqs
10^5 cores
10 hours
1-Frequency
Sim + Map
M3/
GCP
Rate
Year
1.0E+12
1.0E+11
1.0E+10
1.0E+09
1.0E+08
1.0E+07
1.0E+06
1.0E+05
1.0E+04
1.0E+03
1.0E+02
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014


Simulation Capability Over Time
Samples/Second
Samples/Core-Second
1.0E+12
1.0E+11
1.0E+10
1.0E+09
1.0E+08
1.0E+07
1.0E+06
1.0E+05
1.0E+04
1.0E+03
1.0E+02
Rate
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
Year
1-Frequency Sim + Map
9-Frequency Sim + Map
M3/ GCP
TARGET:
10^4 maps
0 freqs
10^5 cores
10 hours

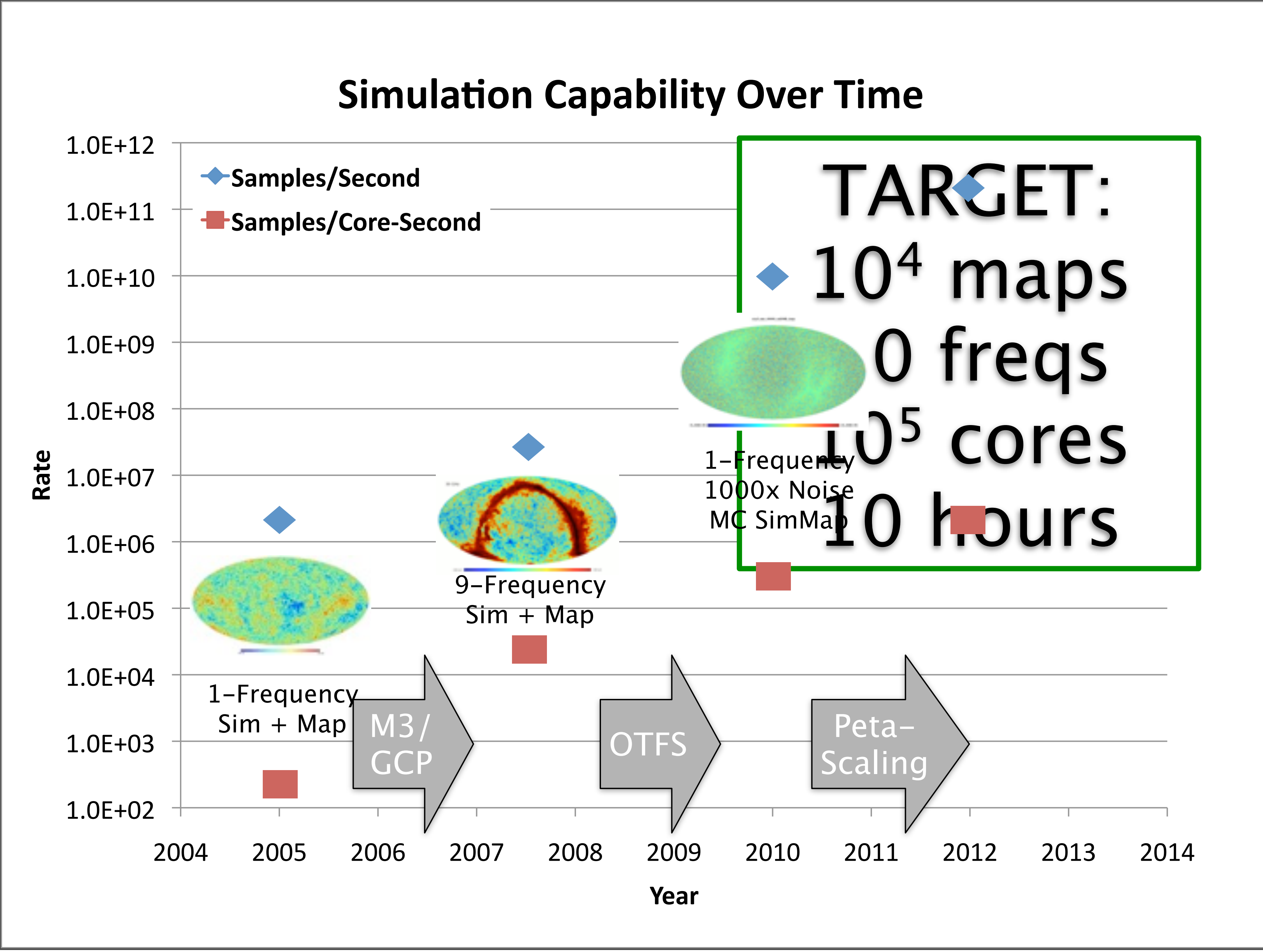

Simulation Capability Over Time
Samples/Second
Samples/Core-Second
1.0E+12
1.0E+11
1.0E+10
1.0E+09
1.0E+08
1.0E+07
1.0E+06
1.0E+05
1.0E+04
1.0E+03
1.0E+02
Rate
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
Year
1-Frequency
Sim + Map
9-Frequency
Sim + Map
M3/
GCP
TARGET:
10^4 maps
10 freqs
10^5 cores
10 hours



Simulation Capability Over Time
Samples/Second
Samples/Core-Second
TARGET:
10^4 maps
10 freqs
10^5 cores
10 hours
9-Frequency
Sim + Map
1-Frequency
Sim + Map
M3/
GCP
OTFS
Rate
Year
1.0E+12
1.0E+11
1.0E+10
1.0E+09
1.0E+08
1.0E+07
1.0E+06
1.0E+05
1.0E+04
1.0E+03
1.0E+02
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014



Thursday, December 16, 2010



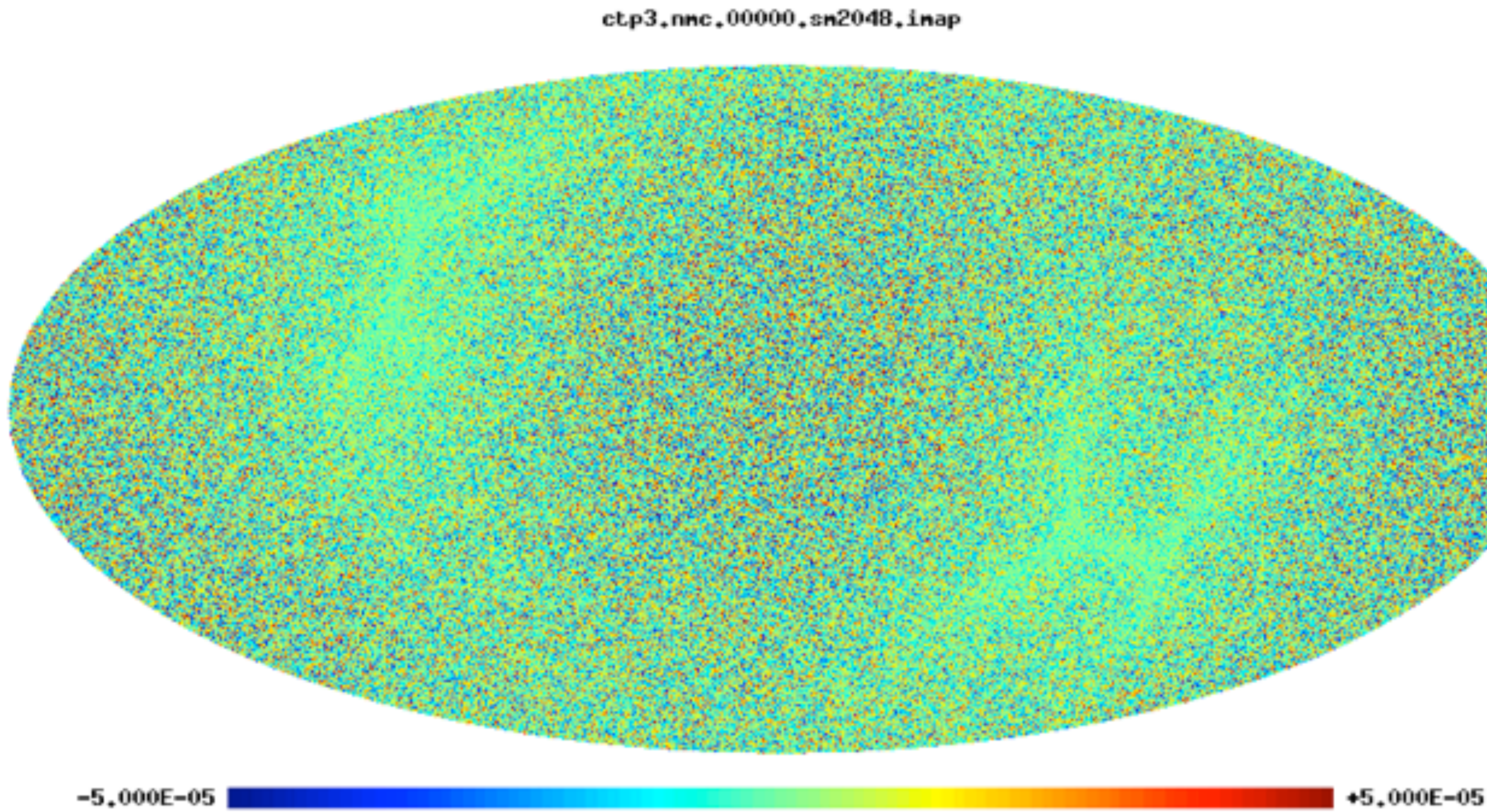
Simulation Capability Over Time
Samples/Second
Samples/Core-Second
Rate
1.0E+12
1.0E+11
1.0E+10
1.0E+09
1.0E+08
1.0E+07
1.0E+06
1.0E+05
1.0E+04
1.0E+03
1.0E+02
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
Year
1-Frequency Sim + Map
9-Frequency Sim + Map
1-Frequency 1000x Noise MC SimMap
M3/ GCP
OTFS
Peta-Scaling
TARGET:
10^4 maps
0 freqs
10^5 cores
10 hours

Current Planck-Scale Simulations

- 1000 x 143 GHz noise maps
  - O(10^{14}) samples, 2TB disk (maps), 2 hours on 20,000

Current Planck-Scale Simulations

- 1000 x 143 GHz noise maps
  - O(10^{14}) samples, 2TB disk (maps), 2 hours on 20,000

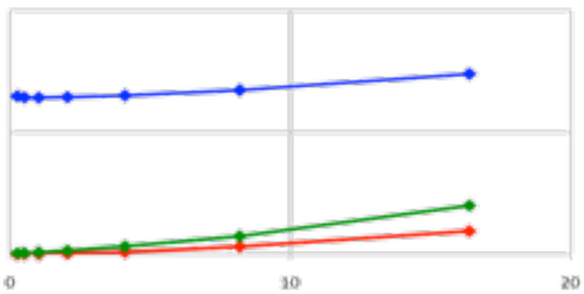


Current Planck-Scale Simulations

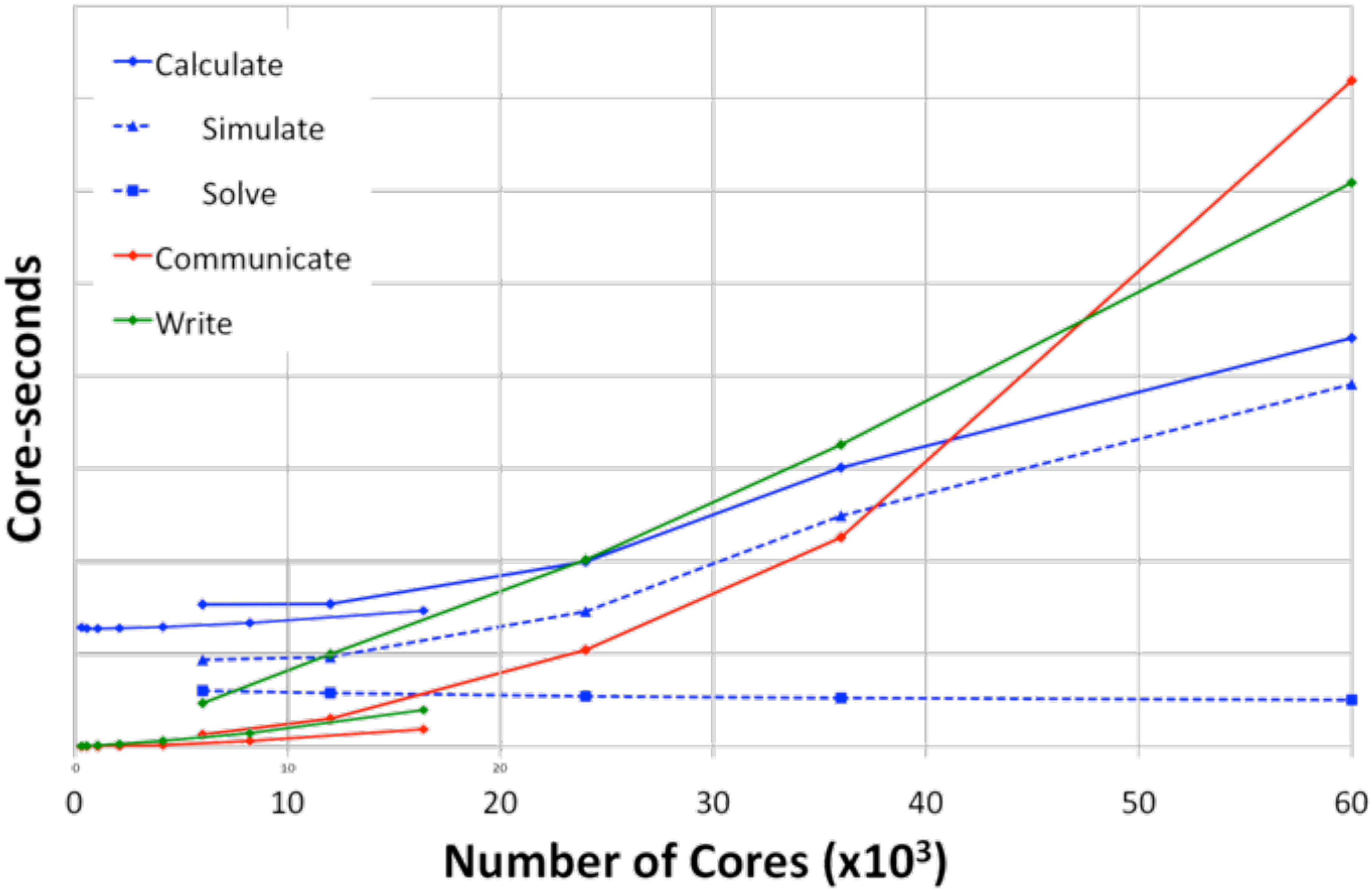
- 1000 x 143 GHz noise maps
  - $O(10^{14})$ samples, 2TB disk (maps), 2 hours on 20,000

Strong Peta-Scaling

Franklin - 100 Tflop/s



Strong Peta-Scaling



The Calculation Bottleneck

- Post-Planck, we're more interested in weak scaling:
  - Calculation scales with N_t; communication & IO with N_p
  - Increasing S/N implies increasing N_t/N_p

    implies increasing calculation/ communication+IO

- Extremely massive heterogeneous systems require new coding paradigms
  - Simple MPI + CPU-per-process won't scale.

- Heterogeneous hierarchy
  - Machine, cabinet, node, processor, die, core, accelerator …
    - Memory, bandwidth, latency hierarchies.
  - Need system-, concurrency- & problem-independent performance
    - Compile- and run-time auto-tuning.

Conclusions

- CMB data set sizes are expected to continue to grow with Moore's Law for the next 15 years, so will continue to be a (b) leading-edge computational challenge for the next 10 M-foldings.

- Good news:
  - Because of our science drivers, we will soon be calculation dominated again (at least until the exa-scale).

- Bad news:
  - Heterogeneity will make it harder to achieve the necessary calculation performance across systems and scales (weak and strong).

- There is no solution, only a process …