Beyond Exascale (?)
- the sky's the limit, or is it sustainable? -

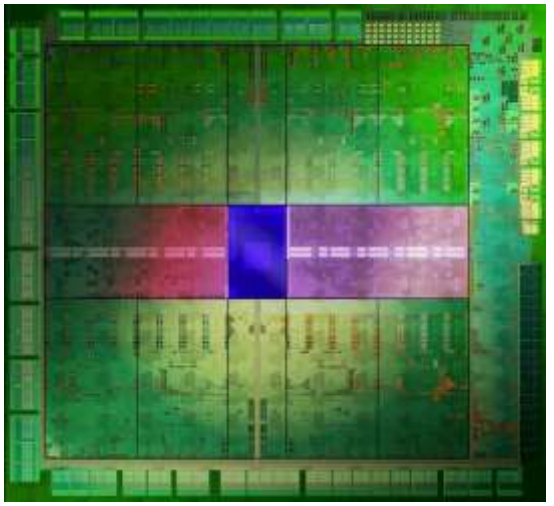
Satoshi Matsuoka

Tokyo Institute of Technology

How much "Flops" will the world produce in 2020?

NVIDIA Tegra K1 (2013)
28nm, 384GFlops SFP
~10W

NVIDIA Tegra 2020
7nm 1TFlop DFP
~10W



2 Billion smartphones/year -> 2 x 10^{21} **or 2 ZetaFlops @ 20 GW (c.f. Entire Japan ~30 GW)**

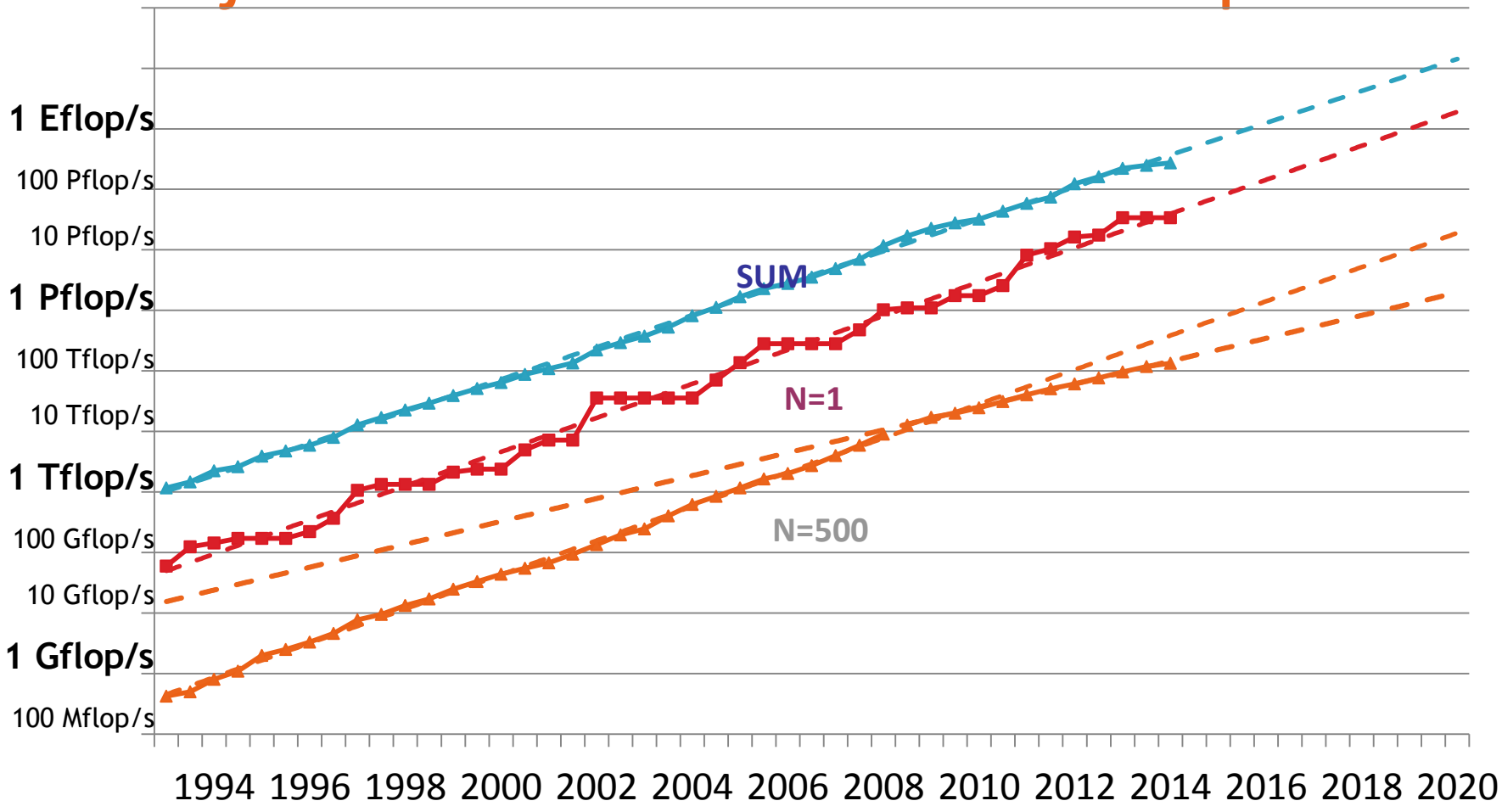
How much energy to drive it?

(Wattage Source Wikipedia)

- Assuming 50GFlops/W
  - Global electricity usage: 2.11 TW-> 105 ZF
  - Global energy usage: 17.1 TW -> 855 ZF
  - Earth solar energy reception: 174 PW-> 610 YF
  - Dyson sphere: 384 YW-> 1.92E37 Flops

  - ***But are we making good use of the capability? (x100 ~= 10 years)***

We are starting to observe our fate: Projected Performance Development

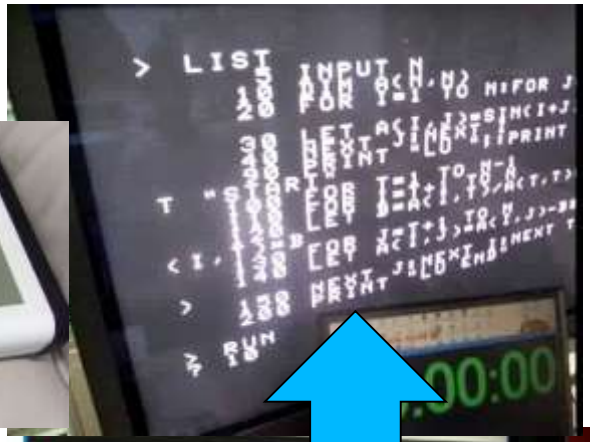


Microprocessor simulation performance circa 1970s

- Hitachi Basic Master (1978)
  - "The first PC in Japan"
  - Motorola 6802--1Mhz, 16KB ROM, 16KB RAM
  - Linpack in BASIC: Approx. 70-80 FLOPS
- We got "simulation" done (in assembly language)
  - Nintendo NES (1982)
    - MOS Technology 6502 1Mhz (Same as Apple II)
  - "Pinball" by Matsuoka & Iwata (now CEO Nintendo)
    - Realtime dynamics + collision + lots of shortcuts
    - Average ~several KFLOPS

Cray-1 (1976)
Linpack
80-90MFlops
(est.)

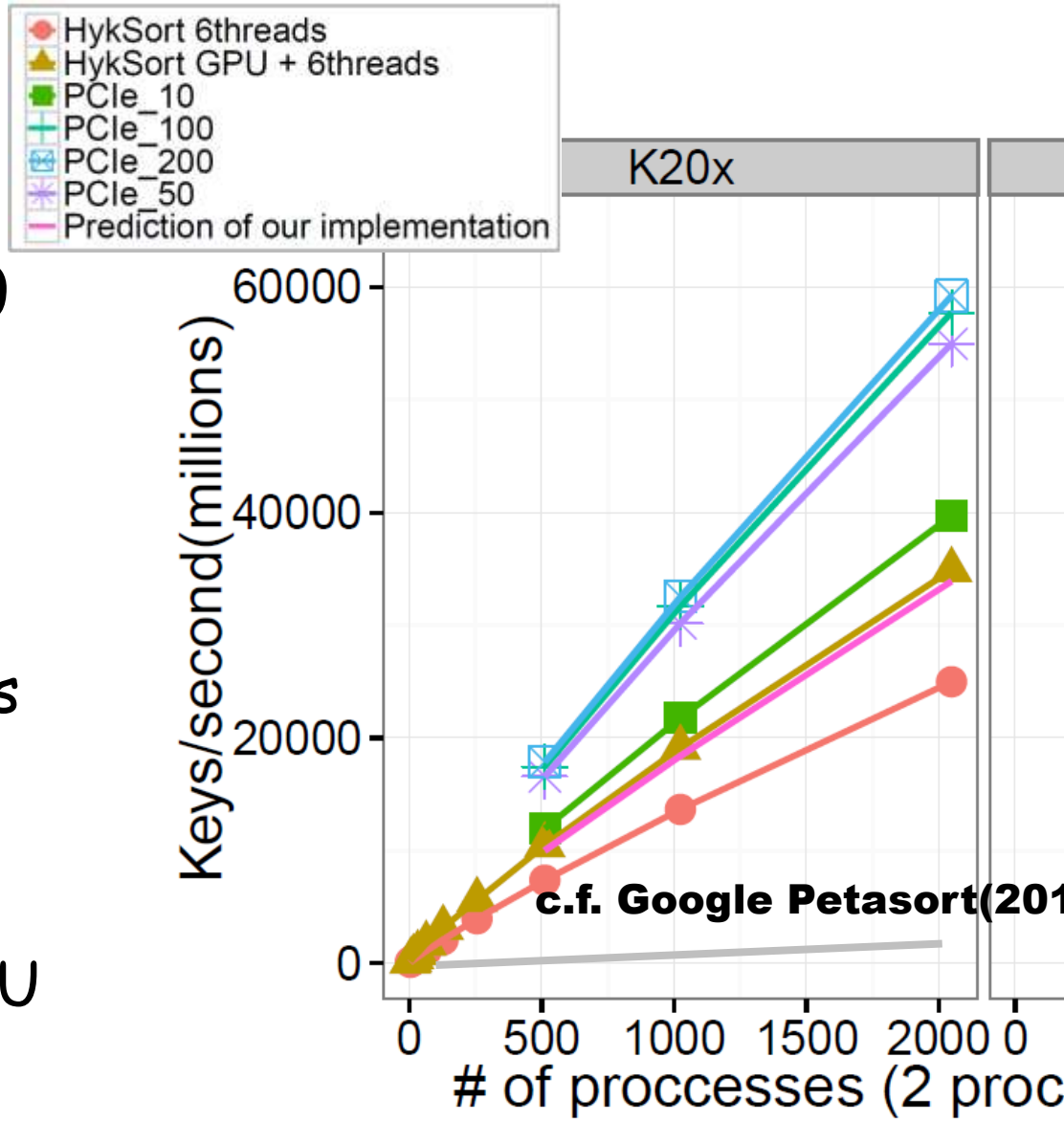
Running Linpack 10



~x100

Where are we now?

- Google Petasort (10 Tera Keys, 100 Byte Records) (MapReduce)
  - 2008: 4K nodes, 8h2m 460M keys/s
  - 2011: 8K nodes, 33min, 5G Keys/s
  - Our on memory GPU sort with NV-link 1K nodes 60G Keys

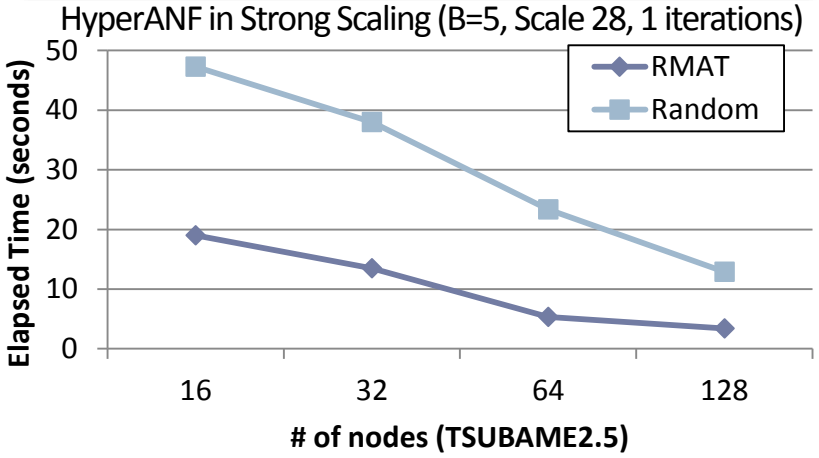


XPregel – X10-based Pregel-like Graph Programming System for convergent architectures

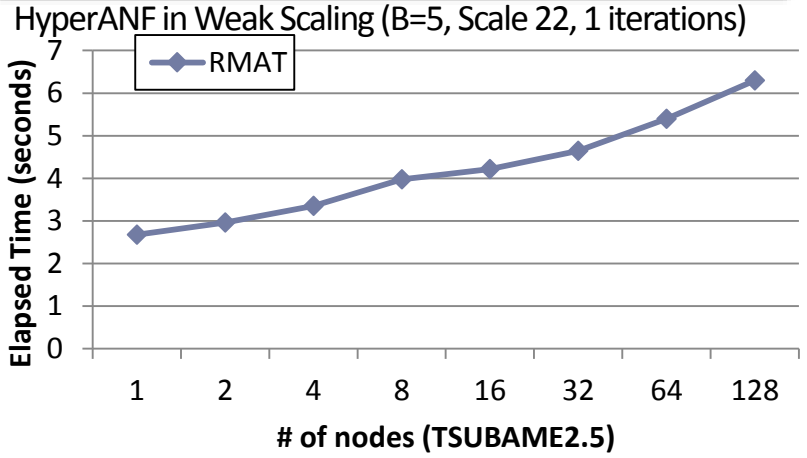
- XPregel optimizations on supercomputers
  1. Utilize MPI collective communication.
  2. Avoid serialization, which enables utilizing fast supercomputer interconnects
  3. Destination of messages computed by a simple bit manipulation thanks to vertex id renumbering.
  4. Optimized message communication when all vertices send the same message to all the neighbor vertices.
  5. Simple API in X10 language.

Performance Evaluation

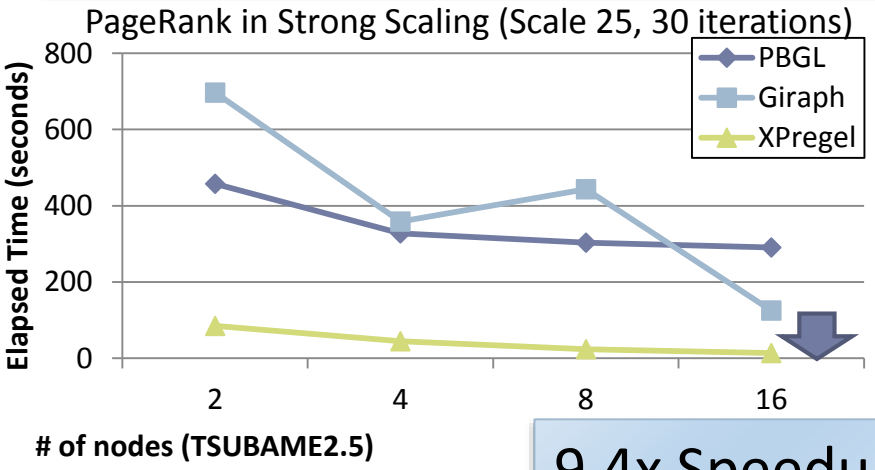
Degree of Separation



Degree of Separation

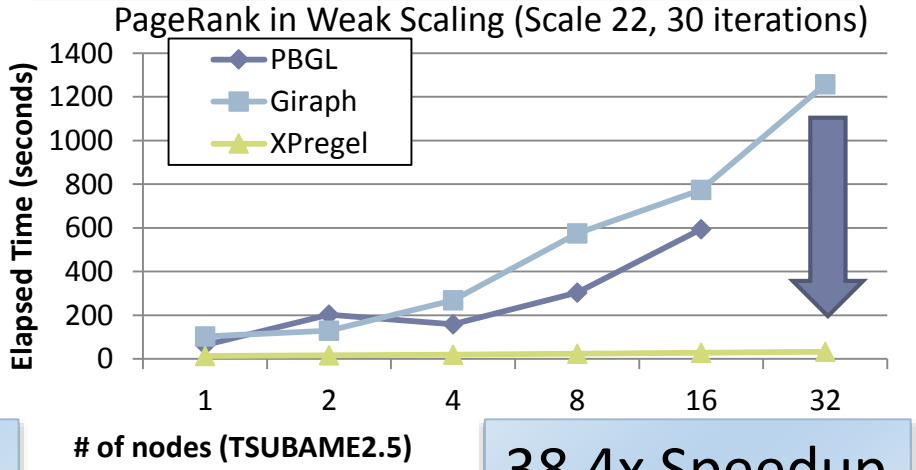


ScaleGraph vs. Giraph, PBGL



9.4x Speedup

ScaleGraph vs. Giraph, PBGL



38.4x Speedup

Hamar (Highly Accelerated Map Reduce) [IEEE Cluster 2014]

- A software framework for large-scale supercomputers w/ many-core accelerators and local NVM devices
  - Abstraction for deepening memory hierarchy
    - Device memory on GPUs, DRAM, Flash devices, etc.
- Features
  - Object-oriented
    - C++-based implementation
    - Easy adaptation to modern commodity many-core accelerator/Flash devices w/ SDKs
      - CUDA, OpenNVM, etc.
  - Weak-scaling over 1000 GPUs
    - TSUBAME2
  - Out-of-core GPU data management
    - Optimized data streaming between device/host memory
    - GPU-based external sorting
  - Optimized data formats for many-core accelerators
    - Similar to JDS format

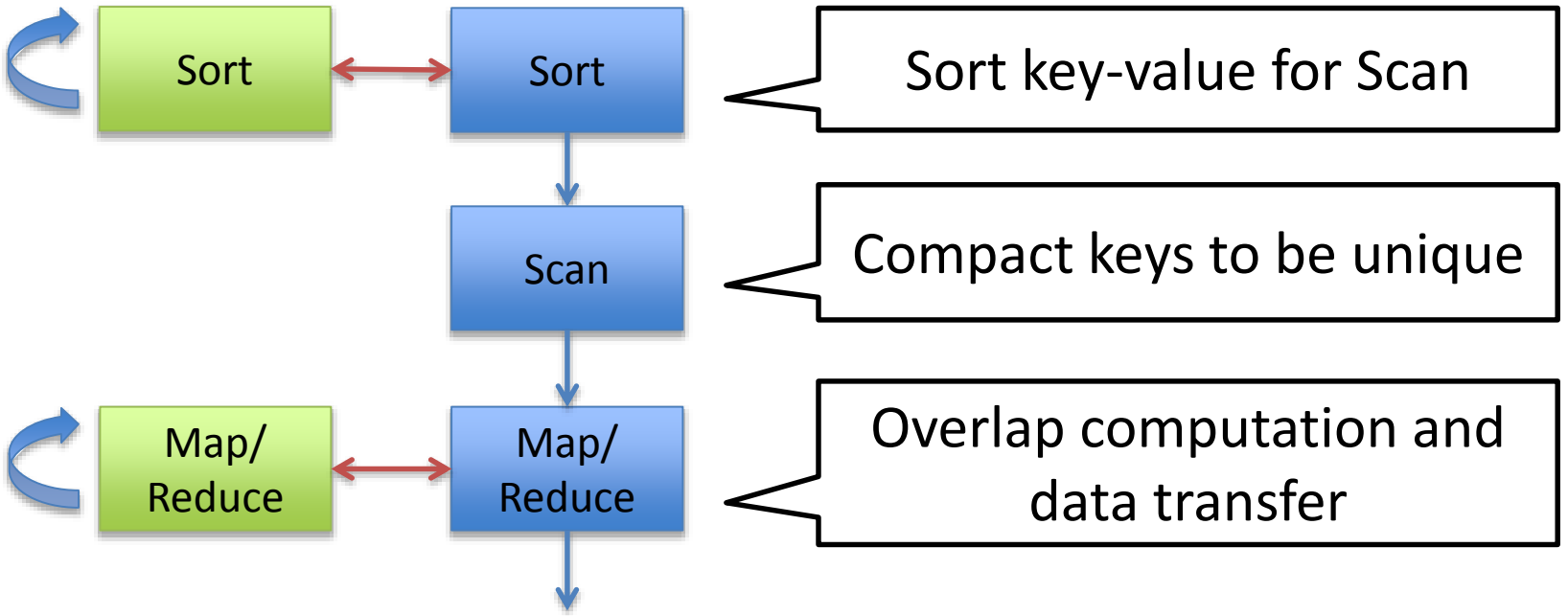






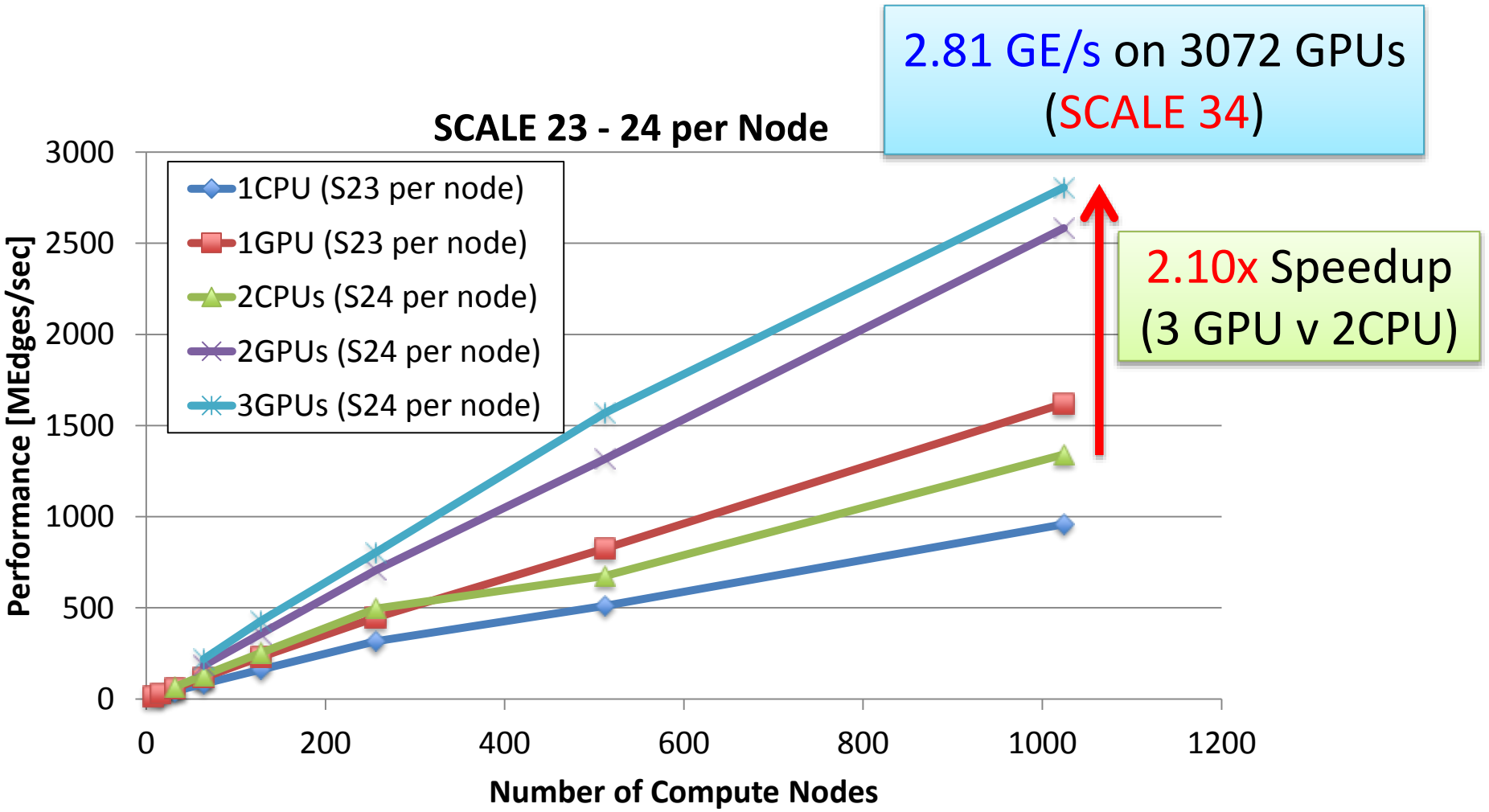
HAMAR Map/Reduce Implementation

- **Optimizations for GPU accelerators**
  - Assign a warp (32 threads) per key for avoiding warp divergence in Map/Reduce
  - Overlapping computation on GPU and data transfer between CPU and GPU
  - Out-of-core GPU Sorting Algorithm



Weak Scaling Performance

- PageRank application on TSUBAME 2.5
- Data size is larger than GPU memory capacity



Third Green Graph 500 List (released June 2014)

In the **Big Data** category:

Rank	MTEPS/W	Site	Machine	G500 rank	Scale	GTEPS	Nodes
1	**59.12**	Kyushu University	GraphCREST-SandybridgeEP-2.4GHz		30	28.48	1
2	**48.29**	Kyushu University	GraphCREST-Sandybridge-EP-2.7GHz		30	31.95	1
3	**35.21**	Tokyo Institute of Technology	GraphCREST-Custom #1		31	13.8	1
4	**28.88**	Tokyo Institute of Technology	MEM-CREST Node #2		30	7.98	1
5	**17.24**	Kyushu University	GraphCREST-Bulldozer		31	13.63	1
6	**14.06**	Tokyo Institute of Technology	TSUBAME-KFC		32	104.31	32
7	**12.48**	The Institute of Statistical Mathematics	ismuv2k2		32	131.43	1
8	**5.41**	Forschungszentrum Julich (FZJ)	JUQUEEN	3	38	5848	16384
9	**4.42**	Argonne National Laboratory	DOE/SC/ANL Mira	2	40	14328	32768

News

June 2014
Third official Green Graph500 list released.

November 2014
Second official Green Graph500 list released.

June 2013
First official Green Graph500 list released.

March 2013
Unofficial Green Graph500 List released.

August 2012
Green Graph 500 Benchmark Code released.

Conclusion

- World could produce Zetaflops of compute – but expensive
- Eventually some limiter will halt our progress
- Wasted cycles are now common with high-level abstractions under the dogma of productivity over performance – however not sustainable
- Better abstractions, or good implementations of them, are necessary for sustainable growth
  - Same as all other industries limited by energy – automotive/transport, construction, manufacturing