

Do You Know What Your I/O Is Doing?

William Gropp

www.cs.illinois.edu/~wgropp



Messages

- Current I/O performance is poor
 - ◆ Even relative to what current systems can achieve
 - ◆ Part of the problem is the I/O interface semantics



Just How Bad Is Current I/O Performance?

- Much of the data (and some slides) taken from “A Multiplatform Study of I/O Behavior on Petascale Supercomputers,” Huong Luu, Marianne Winslett, William Gropp, Robert Ross, Philip Carns, Kevin Harms, Prabhat, Suren Byna, and Yushu Yao, presented at HPDC’15.
 - ◆ This paper has lots more data – consider this presentation a sampling
 - <http://www.hpdc.org/2015/program/slides/luu.pdf>
 - <http://dl.acm.org/citation.cfm?doid=2749246.2749269>
- Thanks to Luu, Behzad, and the Blue Waters staff and project for Blue Waters results
 - ◆ Analysis part of PAID program at Blue Waters



I/O Logs Captured By Darshan, A Lightweight I/O Characterization Tool

- Instruments I/O functions at multiple levels
- Reports key I/O characteristics
- Does not capture text I/O functions
- Low overhead → Automatically deployed on multiple platforms.



Caveats on Darshan Data

- Users can opt out
 - ◆ Not all applications recorded; typically about 1/2 on DOE systems
- Data saved at MPI_Finalize
 - ◆ Applications that don't call MPI_Finalize, e.g., run until time is expired and then restart from the last checkpoint, aren't covered
- About 1/2 of Blue Waters Darshan data not included in analysis
 - ◆ Expect to be fixed soon

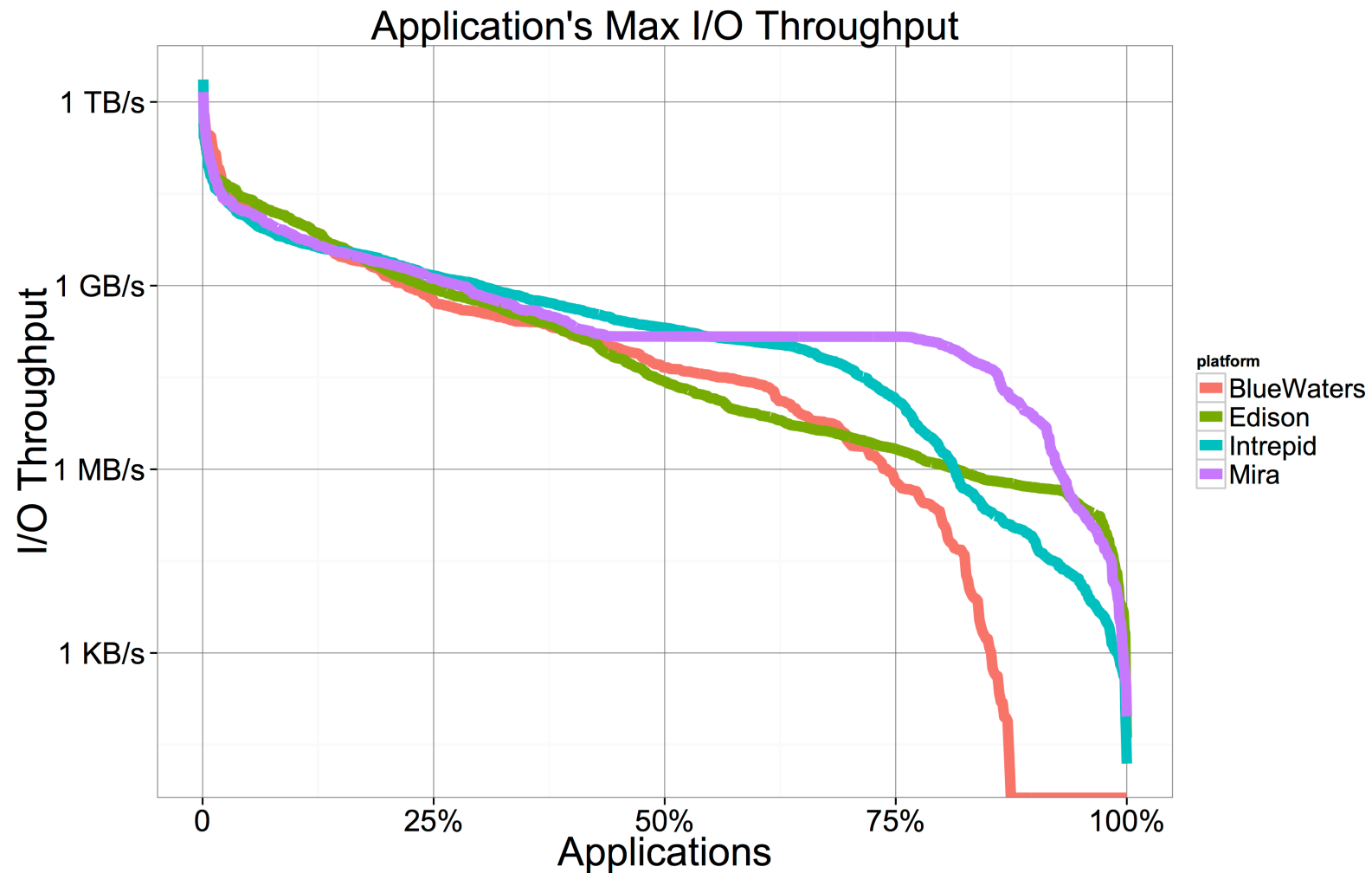


I/O log dataset: 4 platforms, >1M jobs, almost 7 years combined

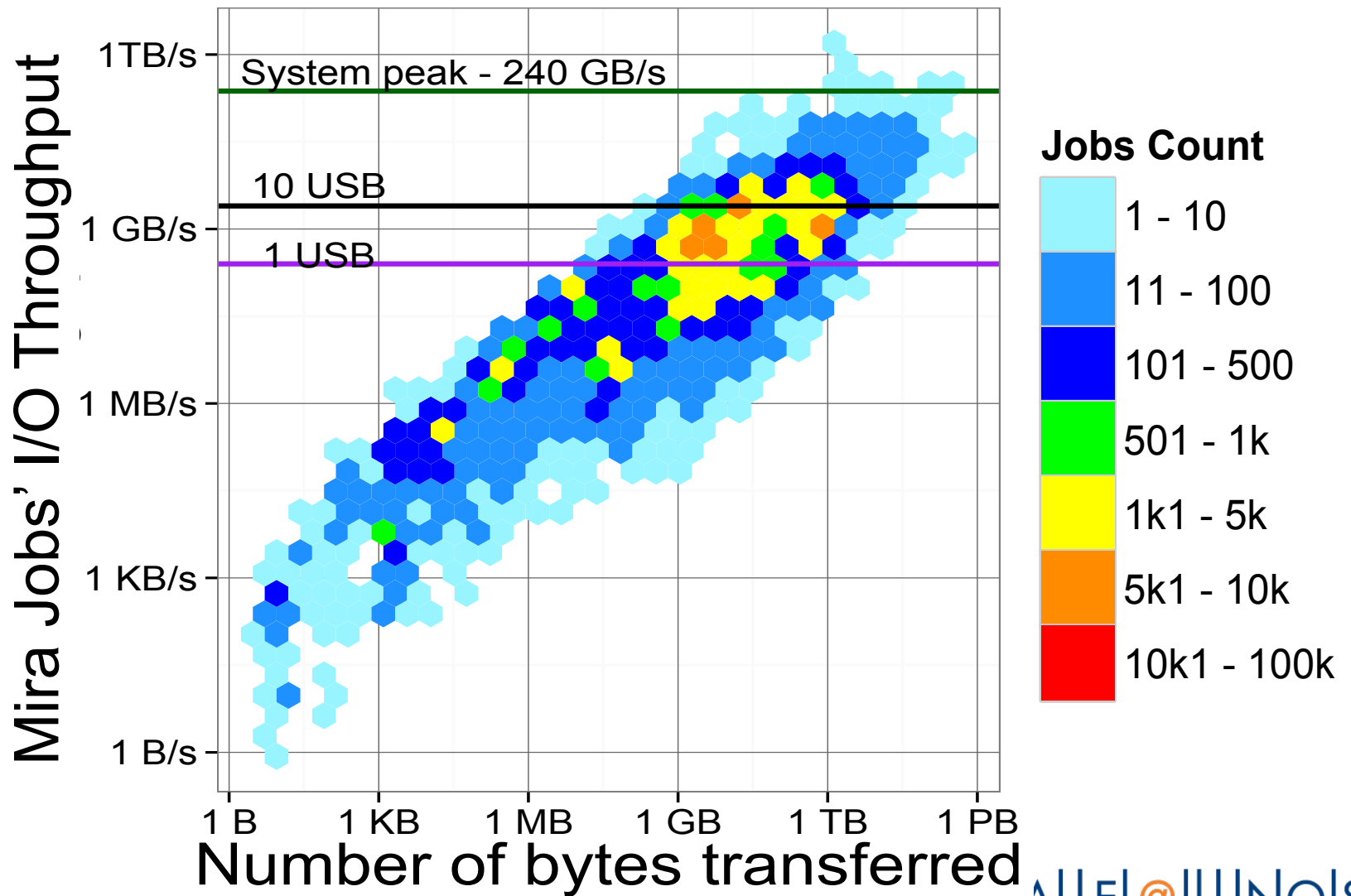
	Intrepid	Mira	Edison	Blue Waters
Architecture	BG/P	BG/Q	Cray XC30	Cray XE6/ XK7
Peak Flops	0.557 PF	10 PF	2.57 PF	13.34 PF
Cores	160K	768K	130K	792K+59K smx
Total Storage	6 PB	24 PB	7.56 PB	26.4 PB
Peak I/O Throughput	88 GB/s	240 GB/s	168 GB/s	963 GB/s
File System	GPFS	GPFS	Lustre	Lustre
# of jobs	239K	137K	703K	300K
Time period	4 years	18 months	9 months	6 months



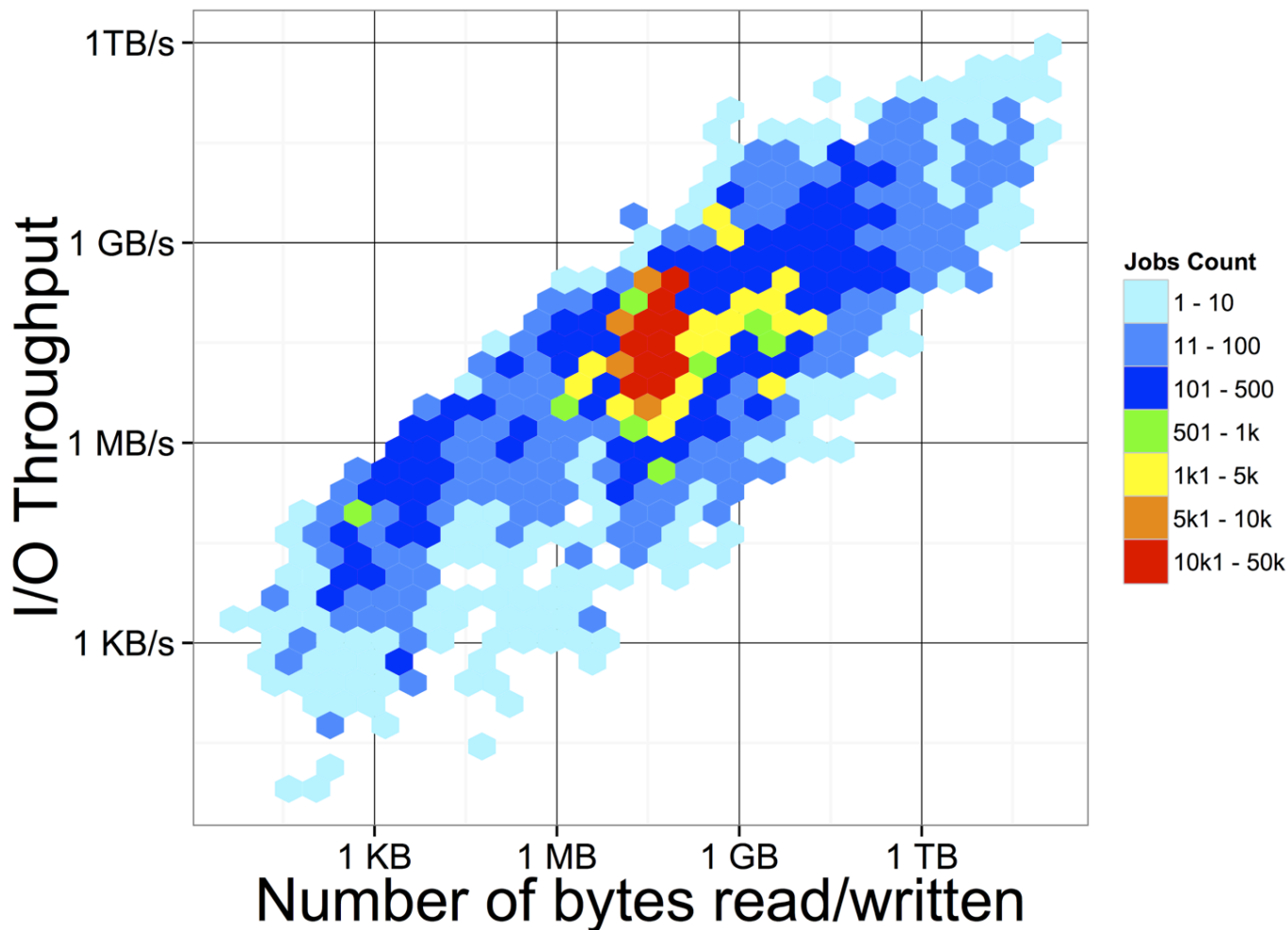
Very Low I/O Throughput Is The Norm



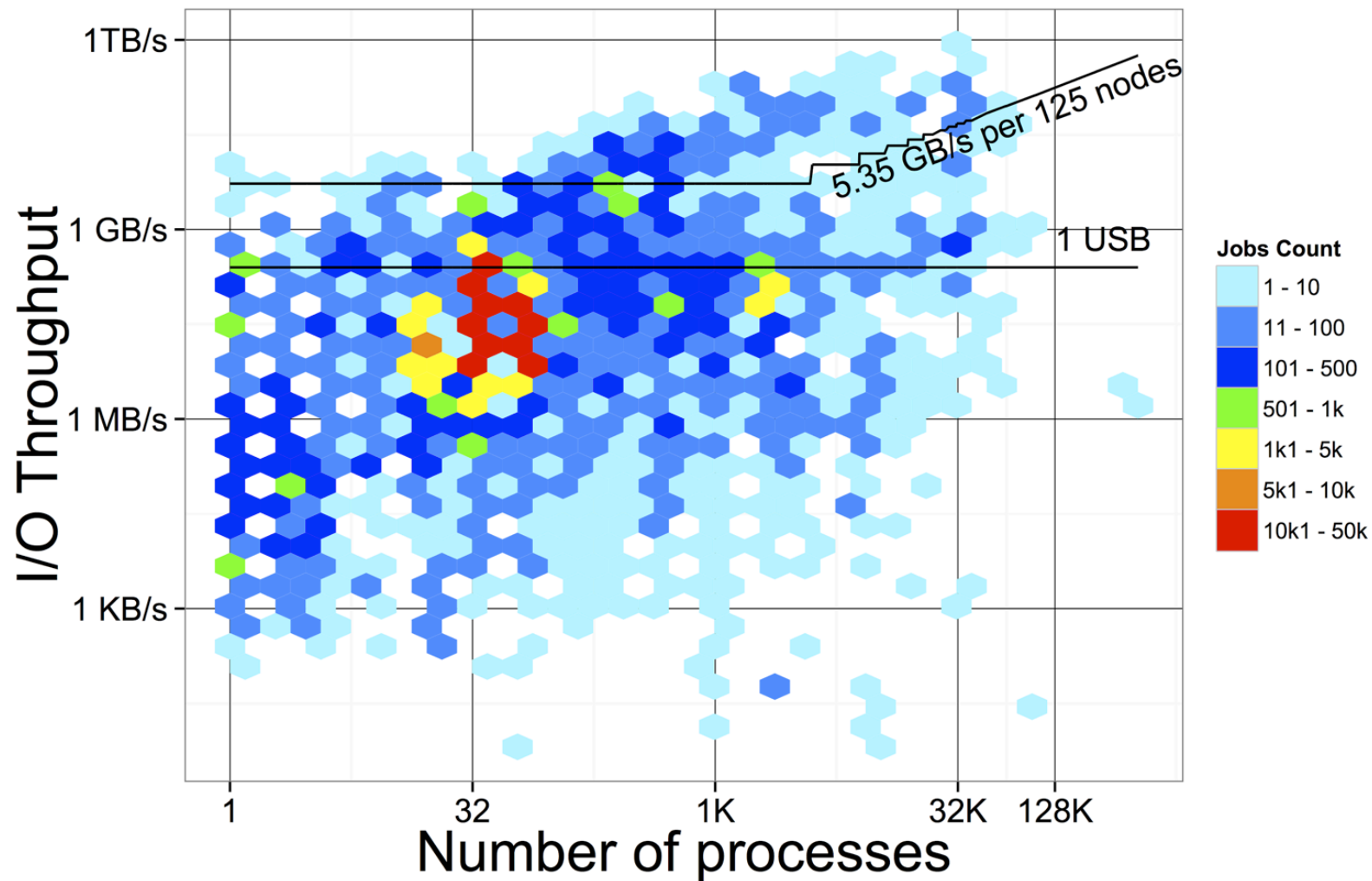
Most jobs transfer little data. Many big-data jobs also have very low thruput

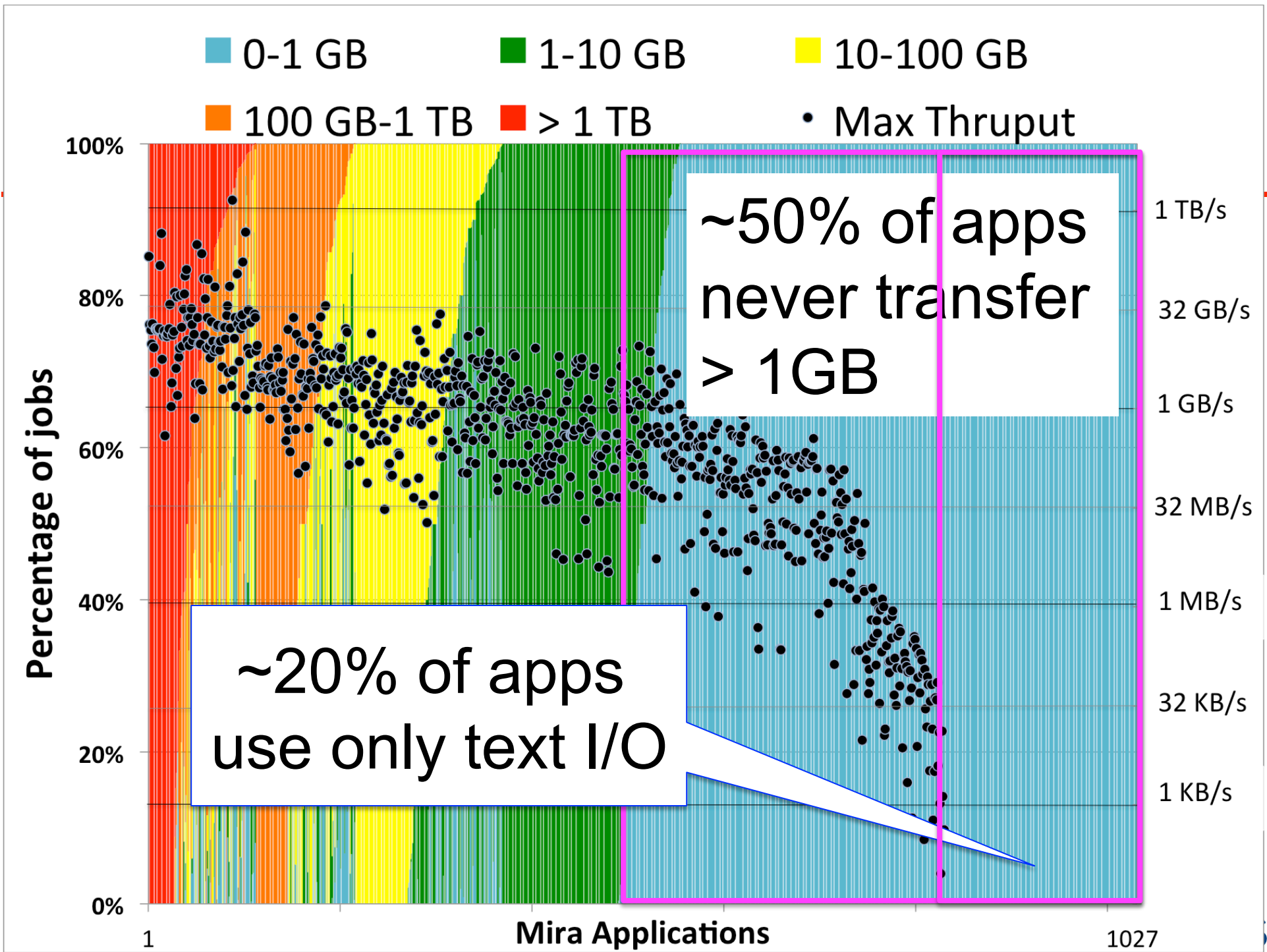


Most Jobs Read/Write Little Data (Blue Waters data)

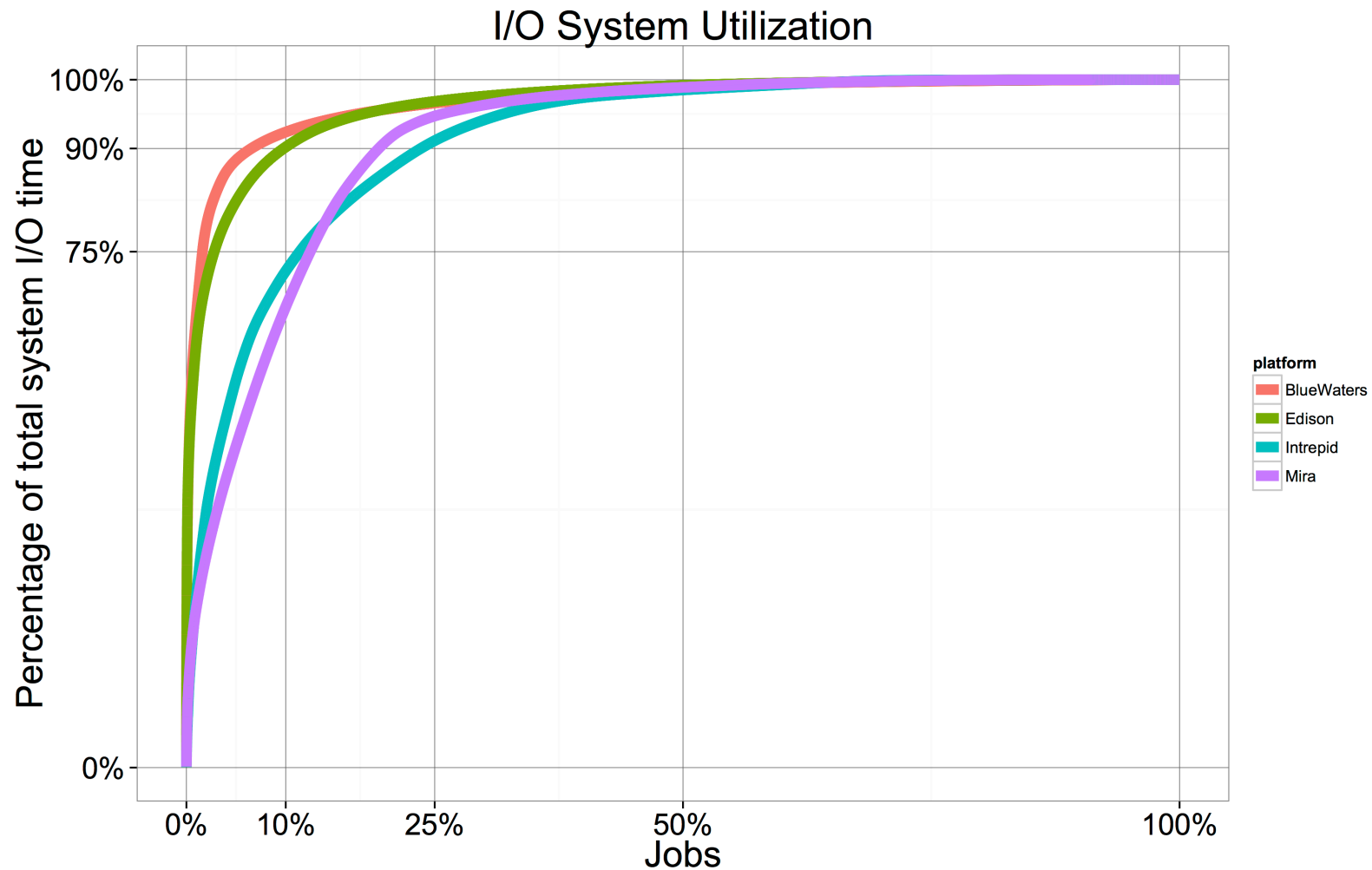


I/O Thruput vs Relative Peak





I/O Time Usage Is Dominated By A Small Number Of Jobs/Apps



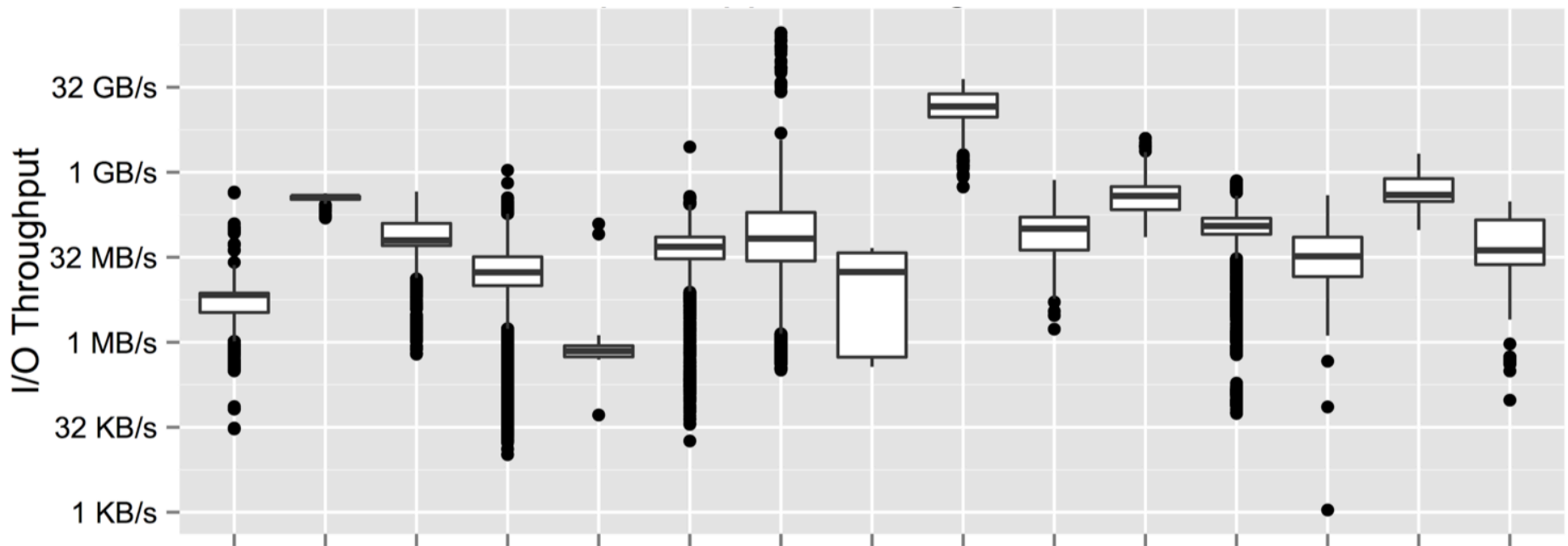
Improving the performance of the top 15 apps can save a lot of I/O time

	Platform I/O time percent	Percent of platform I/O time saved if min thrupt = 1 GB/s
Mira	83%	32%
Intrepid	73%	31%
Edison	70%	60%
Blue Waters	75%	63%

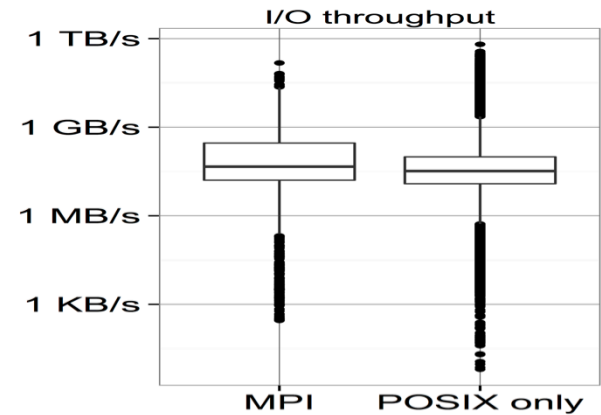
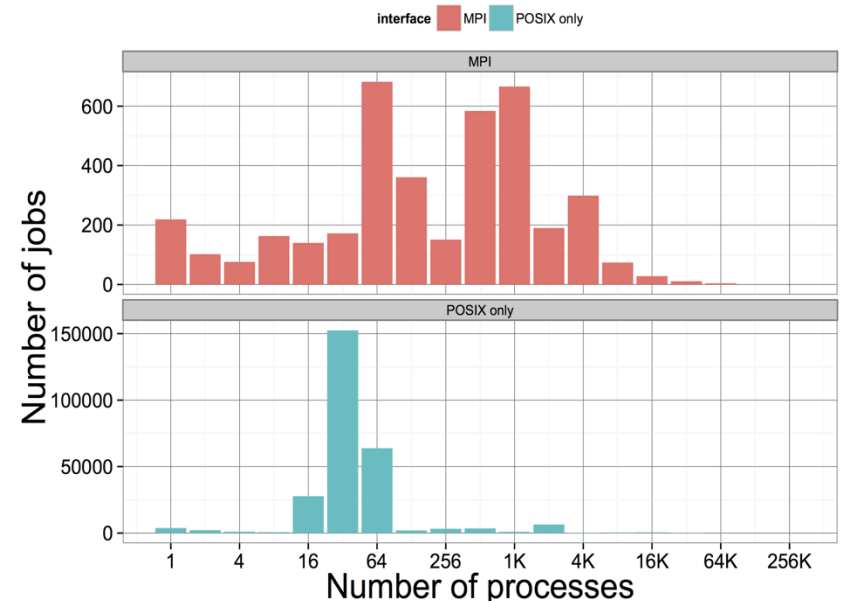
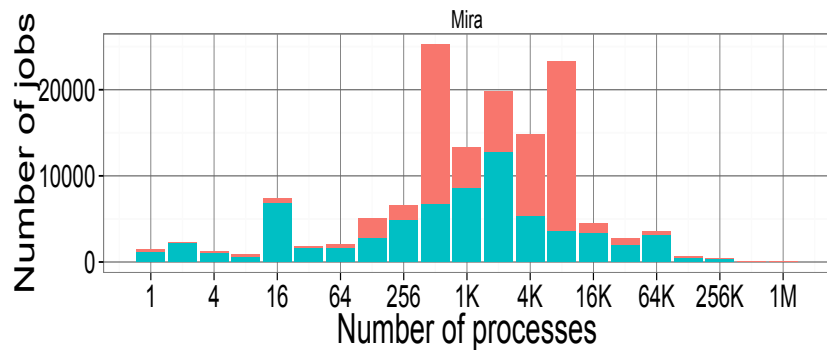
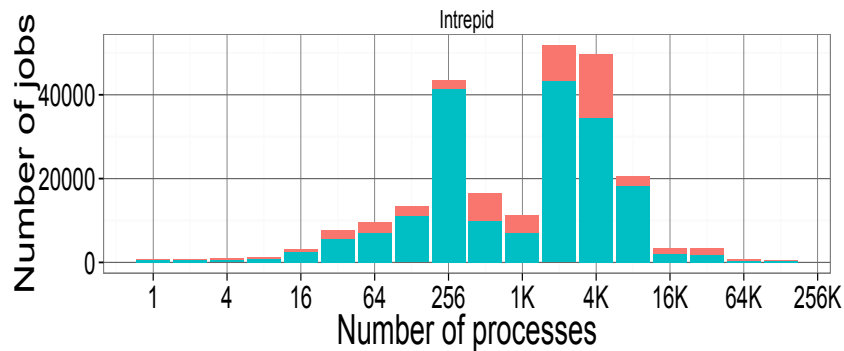
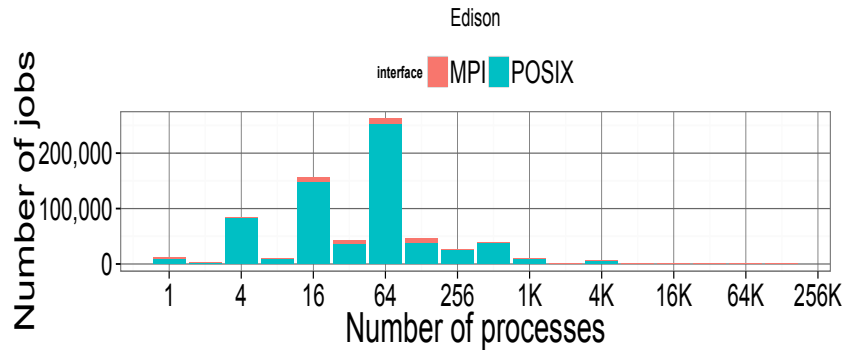


Top 15 apps with largest I/O time (Blue Waters)

- Consumed 1500 hours of I/O time (75% total system I/O time)



POSIX I/O is far more widely used than parallel I/O libraries.



What Are Some of the Problems?

- POSIX I/O has a strong consistency model
 - ◆ Hard to cache effectively
 - ◆ Applications need to transfer block-aligned and sized data to achieve performance
- Files as I/O objects add metadata “choke points”
 - ◆ Serialize operations, even with “independent” files
- Burst buffers will *not* fix these problems – must change the semantics of the operations
- “Big Data” file systems have very different consistency models and metadata structures, designed for their application needs
 - ◆ Why doesn't HPC?
 - There have been some efforts, such as PVFS, but the **requirement** for POSIX has **held up** progress



Summary

- Current I/O performance is poor
 - ◆ Metadata operations often a significant source of poor performance
 - ◆ Related to mismatch between system and user expectations
 - CS Challenge: Better I/O consistency and programming models
 - Math Challenge: Match algorithms to realities of (changing) hardware; need aggregates, realistic model of data transfer costs



Thanks!

- Especially Huong Luu, Babak Behzad
- Funding from:
 - ◆ NSF
 - ◆ Blue Waters
- Partners at ANL, LBNL; DOE funding



PARALLEL@ILLINOIS