TeraByte TokuSampleSort

Bradley C. Kuszmaul
MIT CSAIL, SUNY Stony Brook, Lincoln Laboratory, Tokutek

Jim Gray's Sorting Benchmark Contest

`http://research.microsoft.com/barc/SortBenchmark/`

Sort disk-resident data, & write it to disk.

There are six prizes each year: 3 categories $\times$ 2 classes.

Winners

	Indy	Daytona	Year	Machine
Terabyte	197s	297s	2007	LL TX2500; 400 nodes; 3,600 disks
	435s		2005	80 Itanium2; 2,520 disks
		1,980s	2004	32 Itanium2; 2,320 disks
	1,080s	2,940s	2000	1952 SP; 2,168 disks
	9,060s	9,060s	1998	32-proc Origin 2000
Minute	264GB	214GB	2007	LL TX2500
		40GB	2006	32 Itanium 2; 128 disks
	116GB		2005	80 Itanium 2; 2,520 disks
	32GB	32GB	2004	32 Itanium 2; 2,320 disks
	21.8GB	12GB	2000	
	10.3GB		1999	
	8.41GB	3.5GB	1998	
	3.5GB		1997	
	1.08GB	1.1GB	1995	

I also won the Daytona Penny sort this year on a $300 desktop with 2 disks.
Won 5 out of 6 categories.

The Machine: Lincoln Lab TX2500

A cluster focused on I/O.

- 400 dual-processor servers
- 8GB/server (3.2TB total)
- 6 disk RAID in each node
- Infiniband

The TX2500 is probably around #75 on the Top-500 list.

Terabyte is an In-Core Sort on TX2500

1. Read 1TB from disk into the 3.2TB main memory.
2. Determine which node should receive each record so that node i's records are all $<$ node $i+1$'s records.
3. Send each record to the proper node.
4. Sort in a node.
5. Write to disk.

Terabyte Indy Sort Time breakdown

		Time
Step 0:	Startup	9.7s
Step 1:	Read Data	37.1s
Step 2:	Oversample	0s
Step 3:	Sort sample	44.1s
Step 4:	Calculate Pivots	14.0s
Step 5:	Permute	46.4s
Step 6:	LocalSort	23.9s
Step 7:	Write (& fsync)	11.6s