

Lustre Experience on Cielo

Joel Stevenson (ACES, SNL)

Cielo Lustre L2 Milestone Review, March 11, 2013

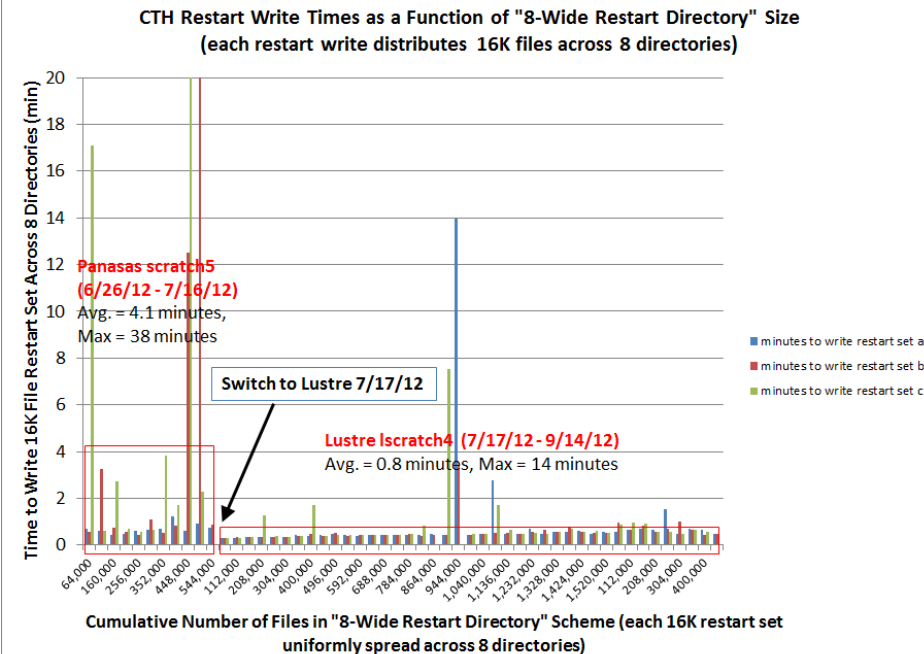
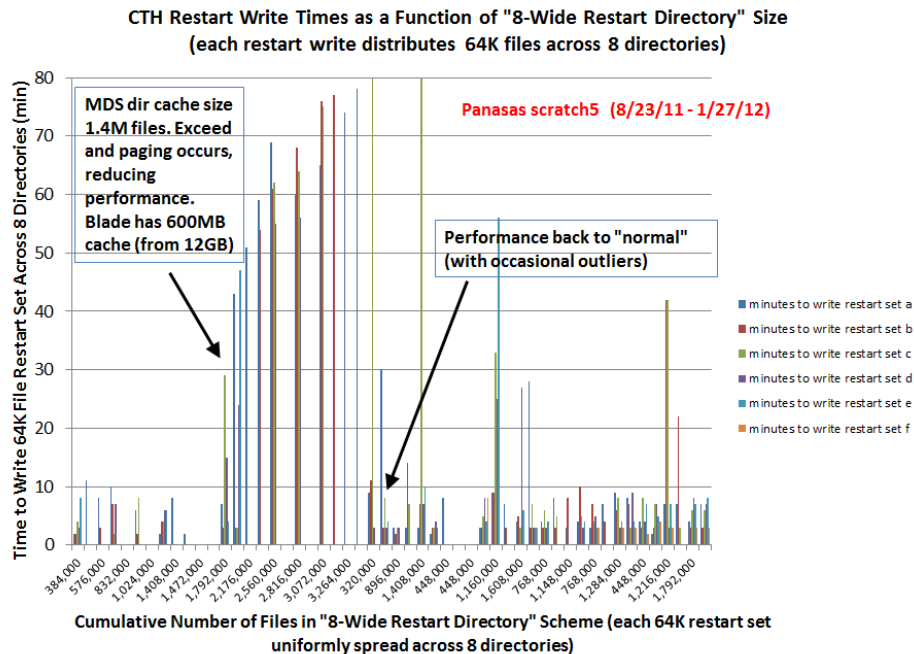
File System Performance with CTH

- Compared write times with Lustre and Panasas for Jason Wilke's CTH simulations (32K core CCC-1, 64K core CCC-2, and 16K core CCC-3)
 - Lustre writes were 1 min; Panasas writes were 1-30 min
 - Simulation progress was faster with Lustre (larger is better in tables below)
 - CCC-1 data set: lustre 310 usec/24 hr (12.9 usec/hr); panasas 251 usec/24hr (10.5 usec/hr)
 - CCC-2 data set: lustre 203 usec/24 hr (8.5 usec/hr); panasas 182 usec/24hr (7.6 usec/hr)
 - CCC-3 data set: lustre 137 usec/24 hr (5.7 usec/hr); panasas 129 usec/24hr (5.4 usec/hr; restart adjust)

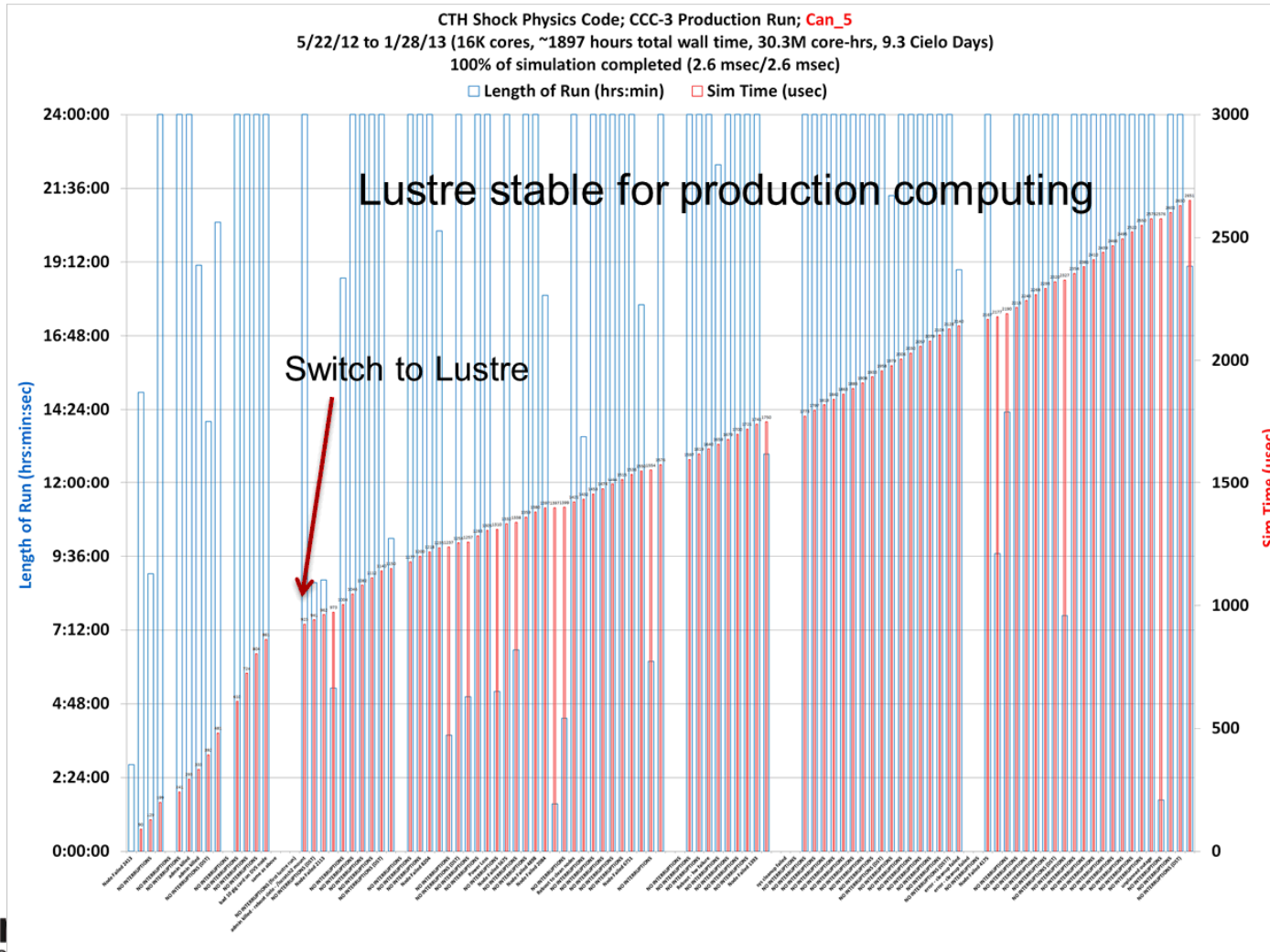
CTH 16.1 with Cray compilers (16.1b from nodegroup1 - update to 7/13/11)									
Date	Time	Job	Problem	# Cores	Max # concurrent I/O processes (Status)	Test files	Sim Time (sec)	write/hr	Run Results
Iscrach4 (Lustre)									
7/2/2012	5:27 PM	434206	28_sage (32K_20-pj1)	32000	2000	32K rest 200GB, 32K spct 10GB, 32K pnt 20GB, 11 jpy	310	12.9	running dedicated - 32K core reservation - iscrach4 at 1% - R: getsize returns 2
7/2/2012	5:29 PM	430956	28_sage (32K_20-pj1)	32000	2000	32K rest 200GB, 32K spct 10GB, 32K pnt 20GB, 11 jpy	310	12.9	running dedicated - 32K core reservation - iscrach4 at 1% - R: getsize returns 2
3/5/2013	2:43 AM	780088	28_sage (32K_20-pj1)	64000	4000	32K rest 200GB, 32K spct 10GB, 32K pnt 20GB, 11 jpy	203	8.5	running dedicated - 64K core reservation - iscrach4 at 1% - R: getsize returns 2
7/4/2012	5:50 AM	435668	28_sage (32K_20-pj1)	16000	4000	32K rest 200GB, 32K spct 10GB, 32K pnt 20GB, 11 jpy	137	5.7	running dedicated - 16K core reservation - iscrach4 at 1% - R: getsize returns 2
scratch4, scratch5 (Panasas PaSca/BB)									
4/5/2011	11:50 AM	58815	28_sage (32K_20-pj1)	32000	2000	32K rest 200GB, 32K spct 10GB, 32K pnt 20GB, 11 jpy	251	10.5	running dedicated - 32K core reservation - iscrach4 at 1% - R: getsize returns 2
8/23/2011	5:50 PM	132743	28_sage (32K_20-pj1)	64000	4000	32K rest 200GB, 32K spct 10GB, 32K pnt 20GB, 11 jpy	182	7.6	running dedicated - 64K core reservation - iscrach4 at 1% - R: getsize returns 2
6/26/2012	8:49 AM	357913	28_sage (32K_20-pj1)	16000	4000	32K rest 200GB, 32K spct 10GB, 32K pnt 20GB, 11 jpy	129	4.3	running dedicated - 16K core reservation - iscrach4 at 1% - R: getsize returns 2

Write Times for CTH Restart Files

- **CCC-2 Panasas**
 - 64K core CCC-2 data set (baton = 4K files)
 - Not overwriting restart files
 - Exceed 1.4M files in “accessed” directories and paging occurs, reducing application performance; stay below threshold and write times = 2-10 minutes
- **CCC-3 Panasas and Lustre**
 - 16K core CCC-3 data set (baton = 4K files)
 - Overwriting restart files
 - Switched to Lustre on 7/17/12
 - Panasas: Avg = 4.1 minutes, Max = 38 minutes
 - Lustre: Avg = 0.8 minutes, Max = 14 minutes



CCC-3 Simulation Progress



Conclusion

- The Cielo file system transition to Lustre has been completed, providing increased performance and reliability needed for full production of the Cielo platform
- The Cielo file system has demonstrated an acceptable production user environment for production capability-class simulations - application users can get their work done on Cielo without major file system issues