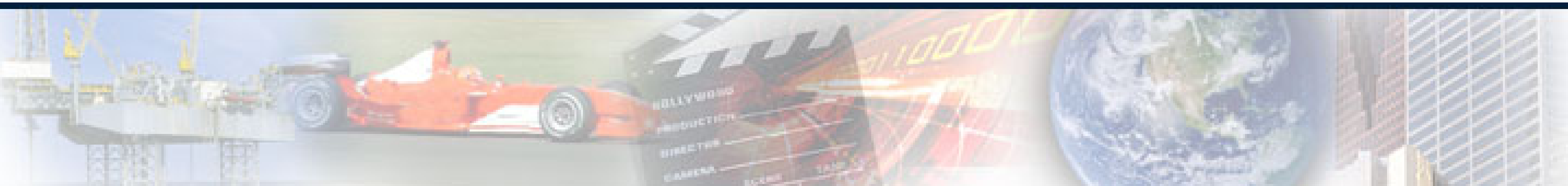
SOS10 Complexity Panel

Garth Gibson
CTO, Panasas, and Assoc. Prof, CMU
ggibson@panasas.com

March 9, 2006

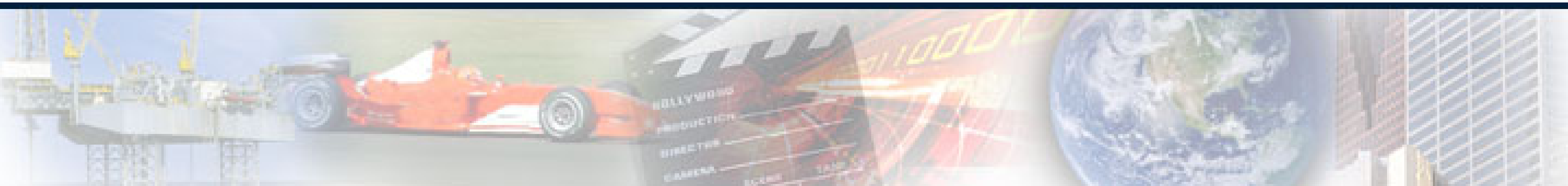


Unable to resist another vehicle analogy

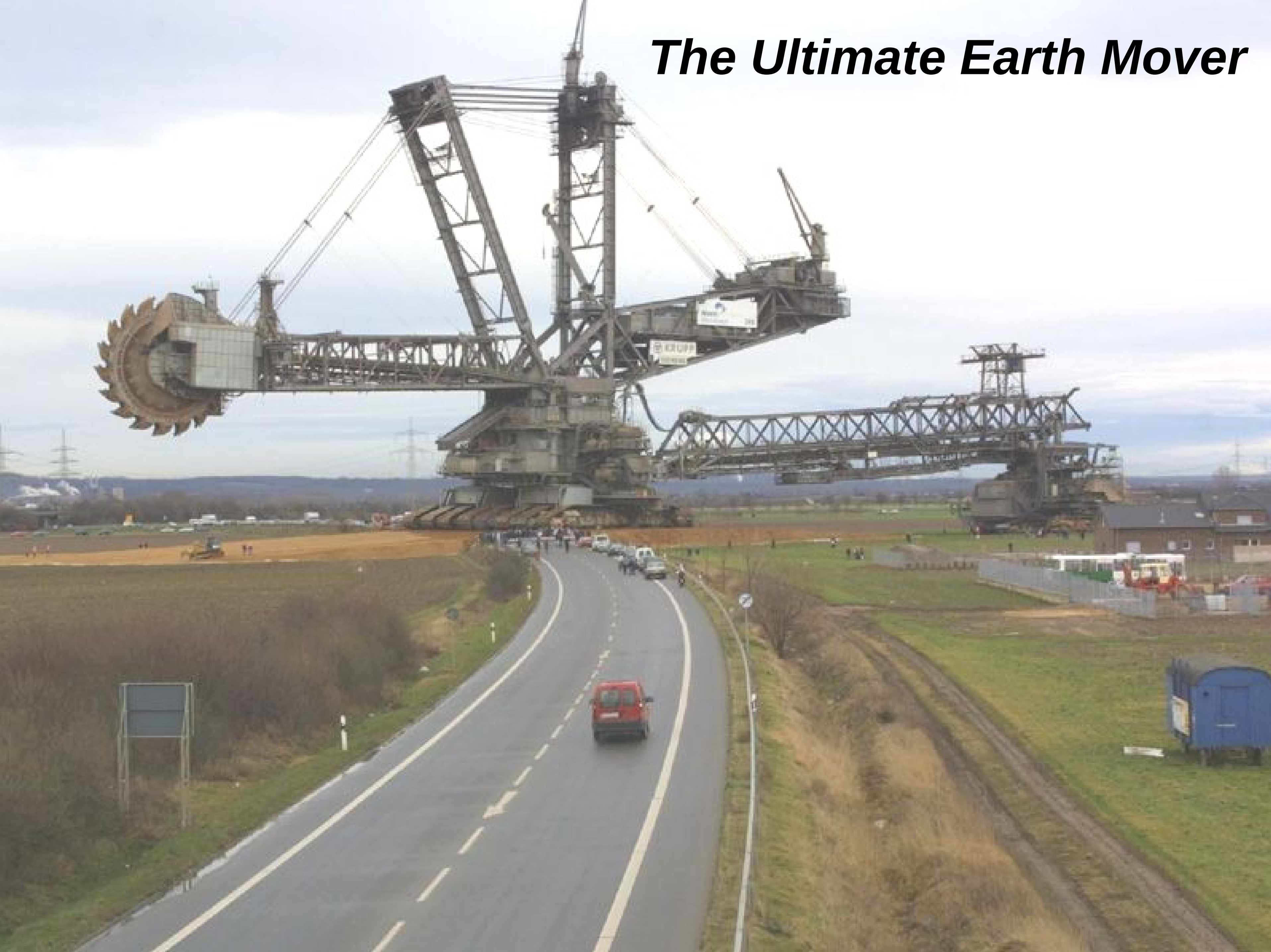
18 wheeler? Blah! Downright ordinary!

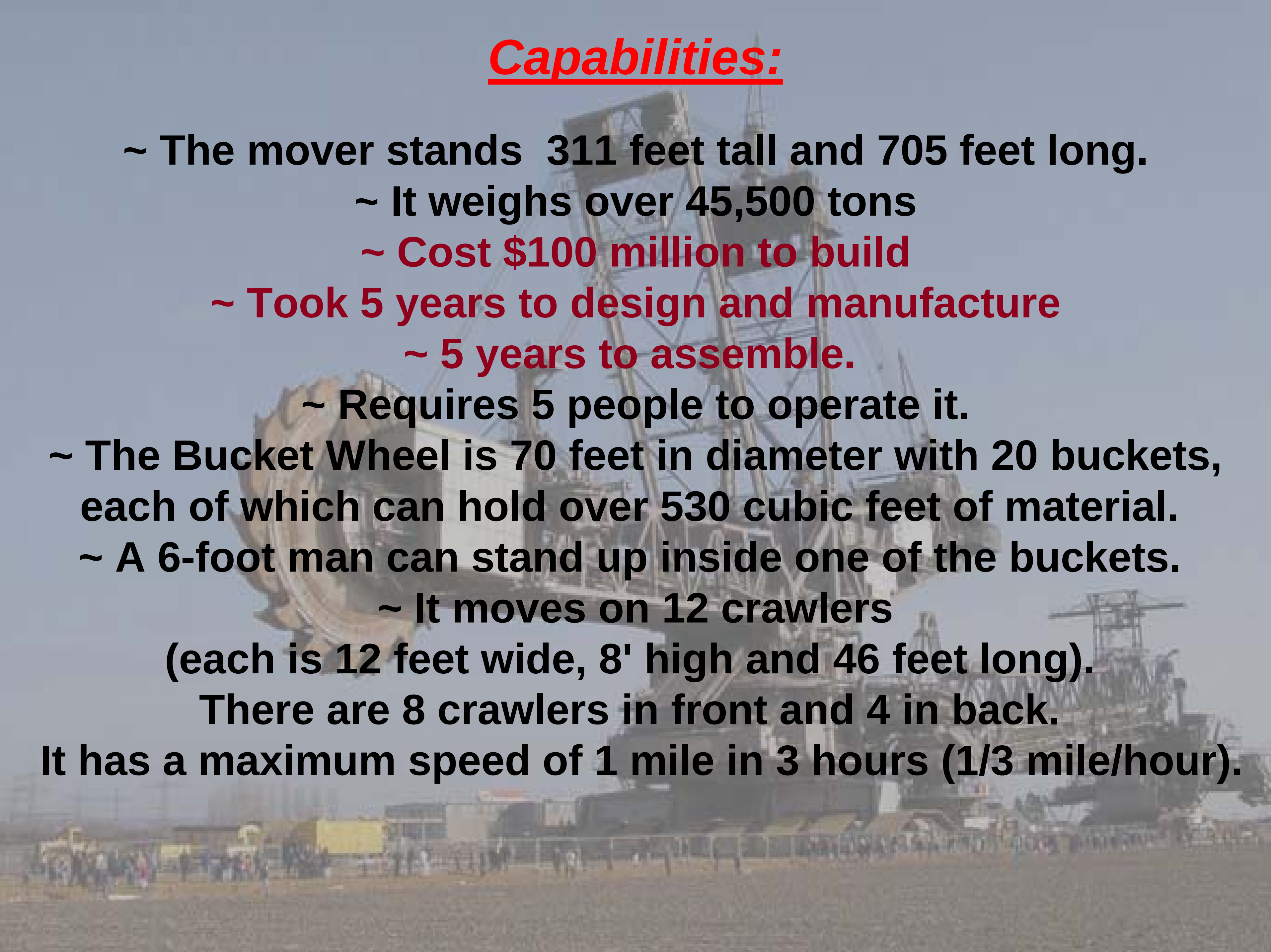
Let me show you a CAPABILITY vehicle!

March 9, 2006



The Ultimate Earth Mover

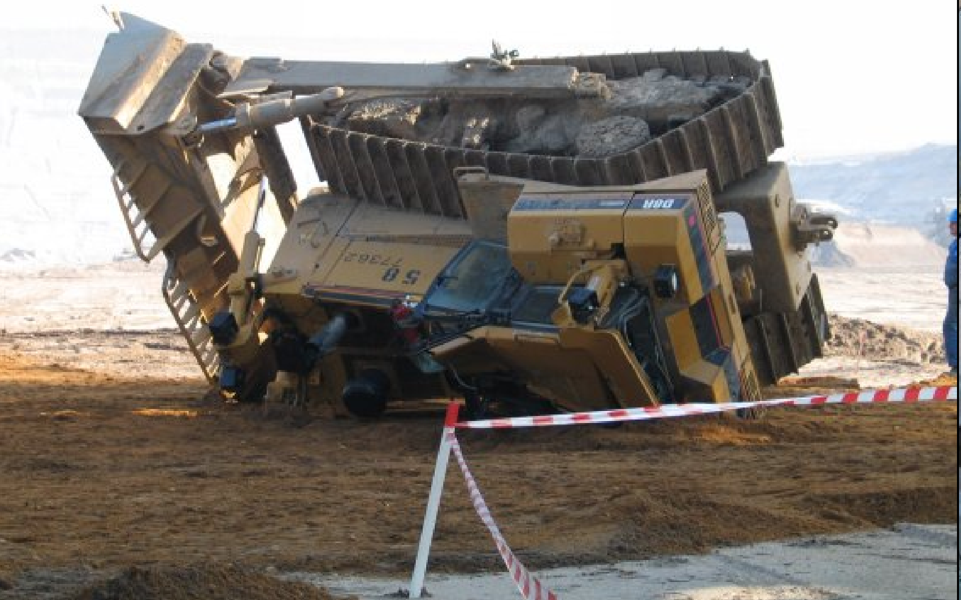




Capabilities:

~ The mover stands 311 feet tall and 705 feet long.
~ It weighs over 45,500 tons
~ Cost $100 million to build
~ Took 5 years to design and manufacture
~ 5 years to assemble.
~ Requires 5 people to operate it.
~ The Bucket Wheel is 70 feet in diameter with 20 buckets, each of which can hold over 530 cubic feet of material.
~ A 6-foot man can stand up inside one of the buckets.
~ It moves on 12 crawlers
(each is 12 feet wide, 8' high and 46 feet long).
There are 8 crawlers in front and 4 in back.
It has a maximum speed of 1 mile in 3 hours (1/3 mile/hour).

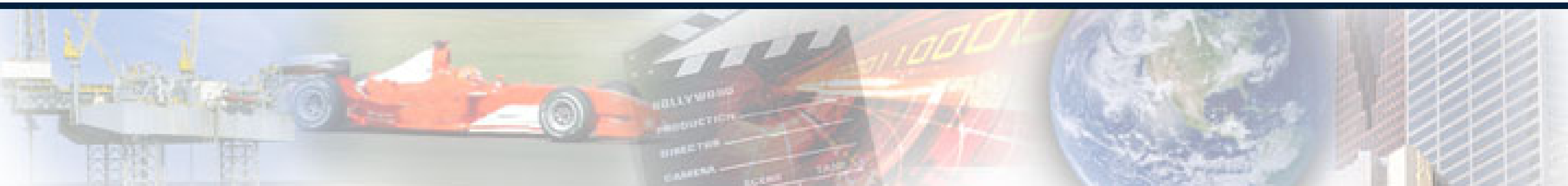
A Capacity Vehicle?



SOS10 Complexity Panel

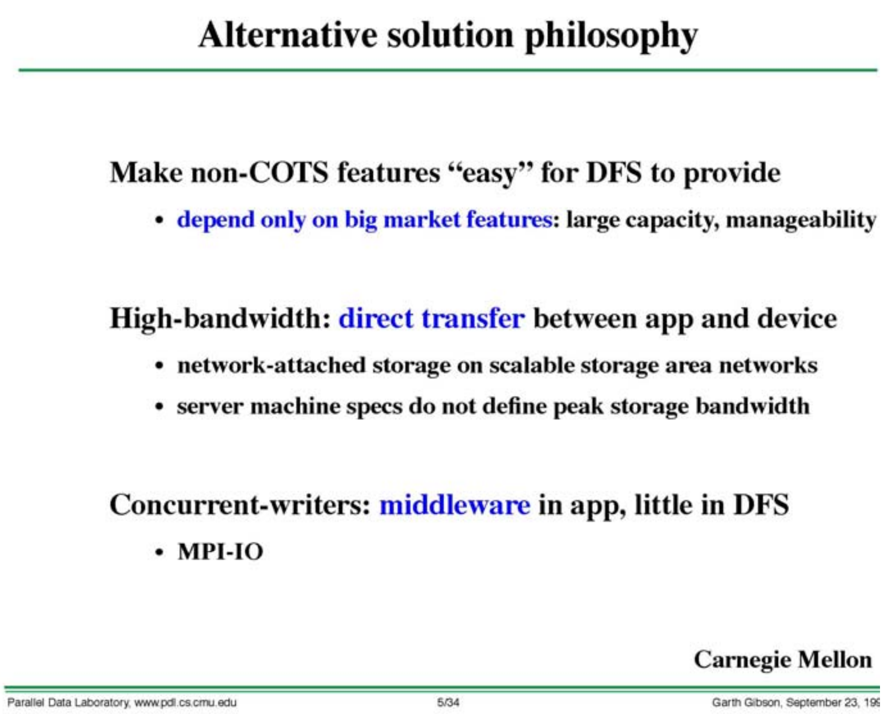
Garth Gibson, Carnegie Mellon University & Panasas Inc.
Harriet Coverston, Sun Microsystems Inc.
Alok Choudhary, Northwestern University
Rob Ross, Argonne National Laboratory
Barney Maccabe, University of New Mexico
Roger Haskin, IBM Almaden Research

March 9, 2006



What have I been doing?

- 6.5 years ago at the Scalable Global Parallel File Systems workshop
- Panasas & Lustre born from CMU’s object storage & AFS/Coda projects
- Primary goals: high bandwidth with high concurrency made easy



Alternative solution philosophy

Make non-COTS features “easy” for DFS to provide

- depend only on big market features: large capacity, manageability

High-bandwidth: direct transfer between app and device

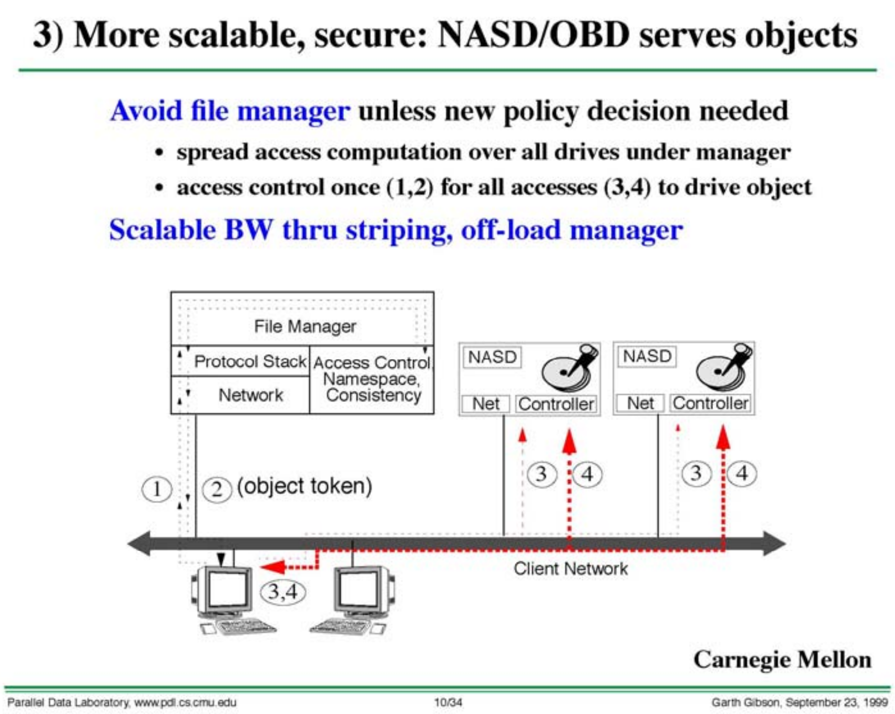
- network-attached storage on scalable storage area networks
- server machine specs do not define peak storage bandwidth

Concurrent-writers: middleware in app, little in DFS

- MPI-IO

Carnegie Mellon

Parallel Data Laboratory, www.pdl.cs.cmu.edu 5/34 Garth Gibson, September 23, 19



3) More scalable, secure: NASD/OBD serves objects

Avoid file manager unless new policy decision needed

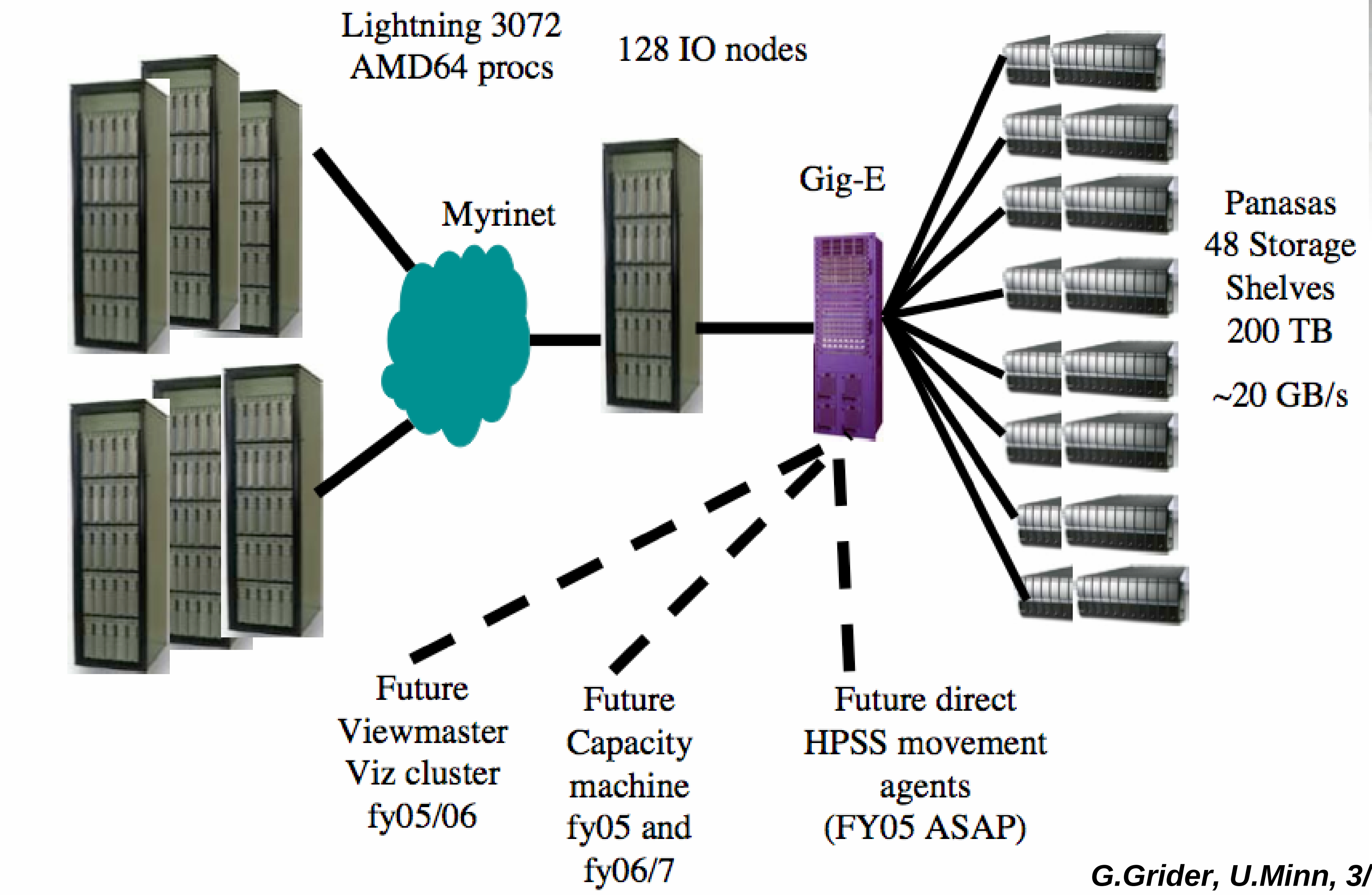
- spread access computation over all drives under manager
- access control once (1,2) for all accesses (3,4) to drive object

Scalable BW thru striping, off-load manager

Carnegie Mellon

Parallel Data Laboratory, www.pdl.cs.cmu.edu 10/34 Garth Gibson, September 23, 1999

PanFS in LANL's Lightning Cluster



G.Grider, U.Minn, 3/05

Also effective for commercial HPC

Petroleum Geo-Services Corporation (PGS)

- Seismic processing outsource company with offices around the world
- Higher performance storage for worldwide seismic processing operations
- Worldwide rollout to 5 continents so far
- High performance for parallel IO in seismic analysis

"The large data sets with which we work require very high bandwidth in order to process data as fast as possible. After evaluating several storage products, none offered the compelling performance and ease-of-management that we receive with Panasas. The Panasas DirectFLOW data path allows us to avoid partitioning the cluster with expensive connections in order to keep up with our heavy bandwidth requirements.

Andy Wrench
DP Computer Systems Manager
PGS Global Computer Resources

Walt Disney Feature Animation

- Creative unit of The Walt Disney Studios producing animated films
- Maximize performance & simplify management
- Thirty Six 5 TB Panasas Storage Cluster shelves (180 TB)
- Over 150,000 operations/sec, 500 MB/s over scalable NFS

QuickTime™ and a
GIF decompressor
are needed to see this picture.

Back to the Panel on Complexity

- Clusters get bigger, applications get bigger, so why would storage getting bigger be any harder?
- Could it be that having every byte of tera- and petabyte stores available to all nodes with good performance for all but minutes a year, when files & volumes are parallel apps on the storage servers, might be a higher standard than compute nodes are held to? (failure…)
- Or perhaps it is deeper and deeper writebehind and readahead, and more and more concurrency, needed to achieve the ever larger contiguous blocks that are needed to minimize seeks in ever wider storage striping. (failure…)
- Or maybe Amdahl's law is hitting us with the need to parallelize more and more of the metadata work which has been serial and synchronous for correctness and error code simplicity in the past. (failure…)
- Or maybe parallel file systems developers have inadequate development tools in comparison to parallel app writers. (test…)
- Or perhaps storage system developers are just wimps. (nerds instead of geeks…)

BANDWIDTH

- 1) In the next decade is the bandwidth transferred into or out of one "high end computing file system"
  - (a) going down 10X or more,
  - (b) staying about the same,
  - (c) going up 10X or more, or
  - (d)"your answer here",
- as a result of the expected increase in computational speed in its client clusters/MPPs, and why?
- Garth (c): 30 GB/s to 1 TB/s is at least 10X
  - But in and of itself this is OK – Object storage scales

SPINDLE COUNT

- 2) In the next decade is the number of magnetic disks in one "high end computing file system"
  - (a) going down 10X or more,
  - (b) staying about the same,
  - (c) going up 10X or more, or
  - (d) "your answer here",
- as a result of the expected increase in computational speed in its client clusters/MPPs, and why?

- Garth (c): 10 year data rate increases (SQRT(MAD))^10
  - This is 8X to 10X based on MAD of 50-60%/yr
  - But if demand goes up 100X, spindle count is still up 10X

CONCURRENCY

- 3) In the next decade is the number of concurrent streams of requests applied to one "high end computing filesystem"
  - (a) going down 10X or more,
  - (b) staying about the same,
  - (c) going up 10X or more, or
  - (d) "your answer here",
- as a result of the expected increase in concurrency in client clusters/MPPs, and why?

- Garth (c): many cores*sockets instead of faster cores
  - Lots more threads, concurrent accesses to storage
  - Seq. data access OK, but metadata concurrency harder

SEEK EFFICIENCY

- 4) In the next decade is the number of bytes moved per magnetic disk seek in one "high end computing file system"
  - (a) going down 10X or more,
  - (b) staying about the same,
  - (c) going up 10X or more, or
  - (d) "your answer here",
- as a result of the expected increase in computational speed in its client clusters/MPPs, and why?

- Garth (b): Possible but not obvious for read/write calls to move more data each, while the cry for 32,000 small file creates/sec means lots more tiny writes
  - Mechanical positioning may continue to hurt big time
  - But file systems still may be faster than DBs for this :-(

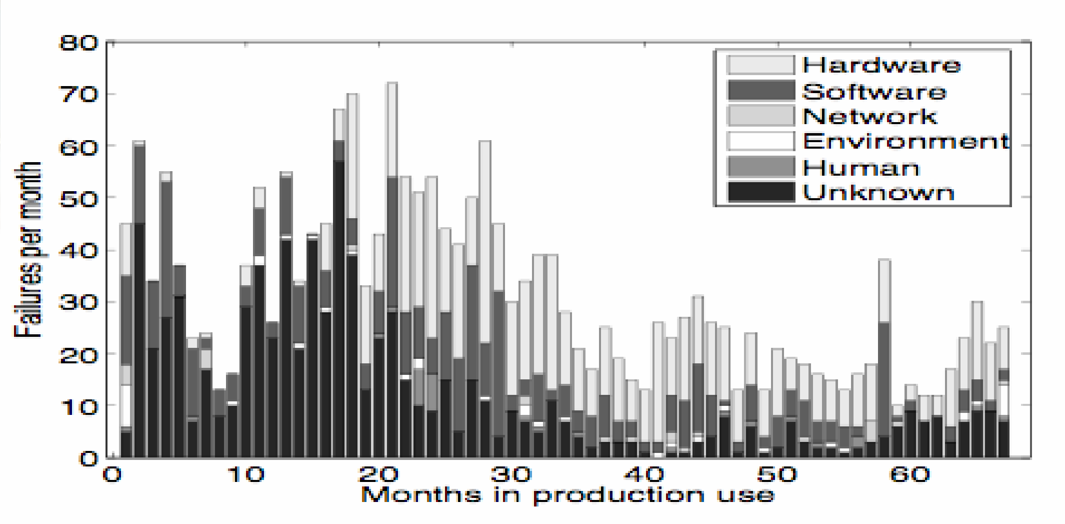
FAILURES

- 5) In the next decade is the number of independent failure domains in one "high end computing file system"
  - (a) going down 10X or more,
  - (b) staying about the same,
  - (c) going up 10X or more, or
  - (d)"your answer here",
- and why?

- Garth (c): as a direct result of all those spindles and and cables
  - All the hard problems come down to the failure cases
  - An now for some interesting data ……

9 YEARS of LANL HEC FAULT DATA

- Failure characteristics differ system to system in rates, causes, and are not stationary over time
- Virtual no widely shared hard data on how HEC computers fail



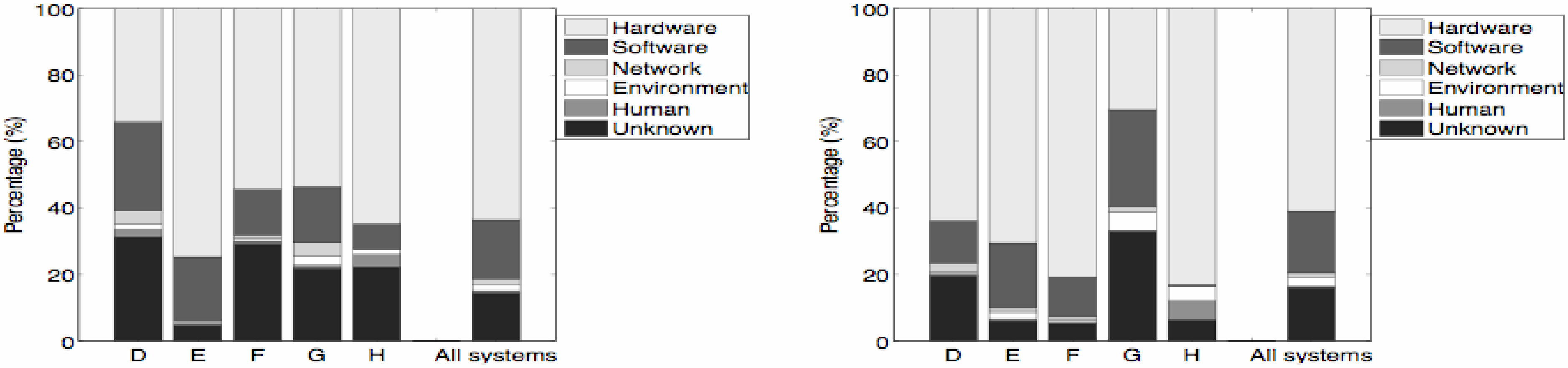


Figure 1: *The breakdown of failures into root causes (left) and the breakdown of downtime into root causes*

COPING WITH COMPLEXITY

- 6) If you have answered (c) one or more times,
  - please explain why these large increases are not going to increase the complexity of storage software significantly?
  - Are you relying on the development of any currently insufficient technologies, and if so, which?

- Garth: Storage developers are at risk here
  - Scaling BW I think we can do
  - Doing that without loss of 9s is hard
  - But scaling metadata rates w/ POSIX consistency is hard
  - Interesting technology: Autonomics, for tuning/healing
  - Interesting technology: Model checking, for protocol correctness

DEVELOPMENT TIME TRENDS

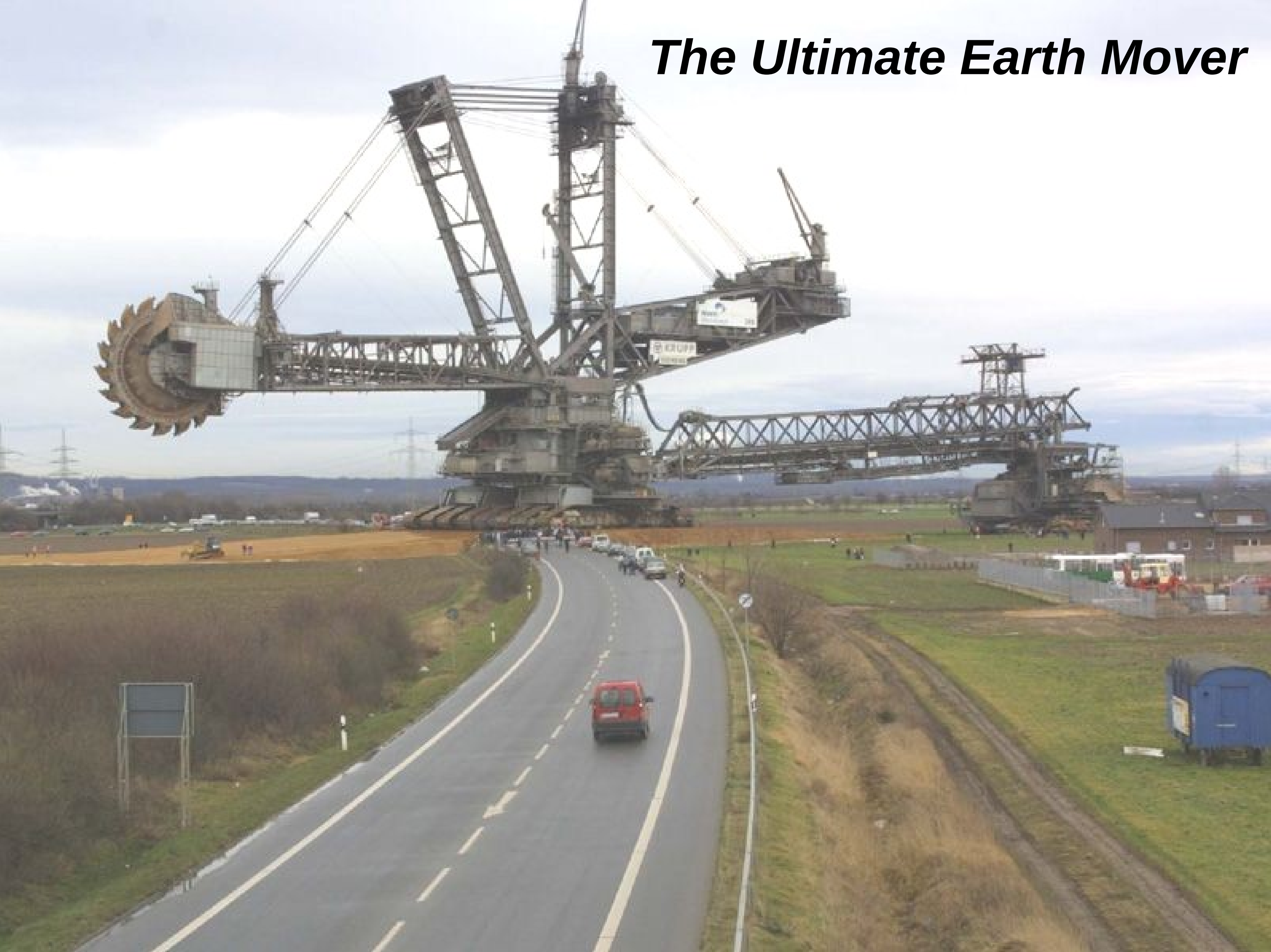
- 7) If complexity is increasing in high end computing file systems, is the time and effort required to achieve acceptable 9s of availability at speed
  - (a) going down 10X or more,
  - (b) staying about the same,
  - (c) going up 10X or more, or
  - (d) "your answer here",
- and why? Are you relying on the development of any currently insufficient technologies, and if so, which?

- Garth (b-c): Can’t face 10X up, but it is increasing
  - Testing can be a big drag with rapidly changing OS/platform
  - To repeat: model checking is interesting

Complexity Questions Summary

	Garth	Harriett	Alok “user”	Rob	Barney FATMagic	Roger Faithbased
1. BW	>10X	>10X	<10X	Up	>10X	>10X
2. Spindles	>10X	>10X	dc	Up		>10X
3. Concurrency	>10X	>10X	>10X	neoPOSIX	>10X	>10X
4. Seek Effic.	~1X	smaller	hidden		?	smaller
5. Failures	>10X	>“size”	HIDE	Up	>10X	<10X
6. Complexity	At risk	STDs	Layers!	Reuse	Users	bugs
7. Dev Time	At risk Model checking	STDs T10 pNFS	Layers & I/F ineff.	3-4 yrs needs to go down	Not allowed: .5M LOC in clutch	Not MDS COTS parts & bugs

The Ultimate Earth Mover



panasas

Next Generation Network Storage

Garth Gibson
ggibson@panasas.com

March 9, 2006