

# Accelerating Workflows with SGI

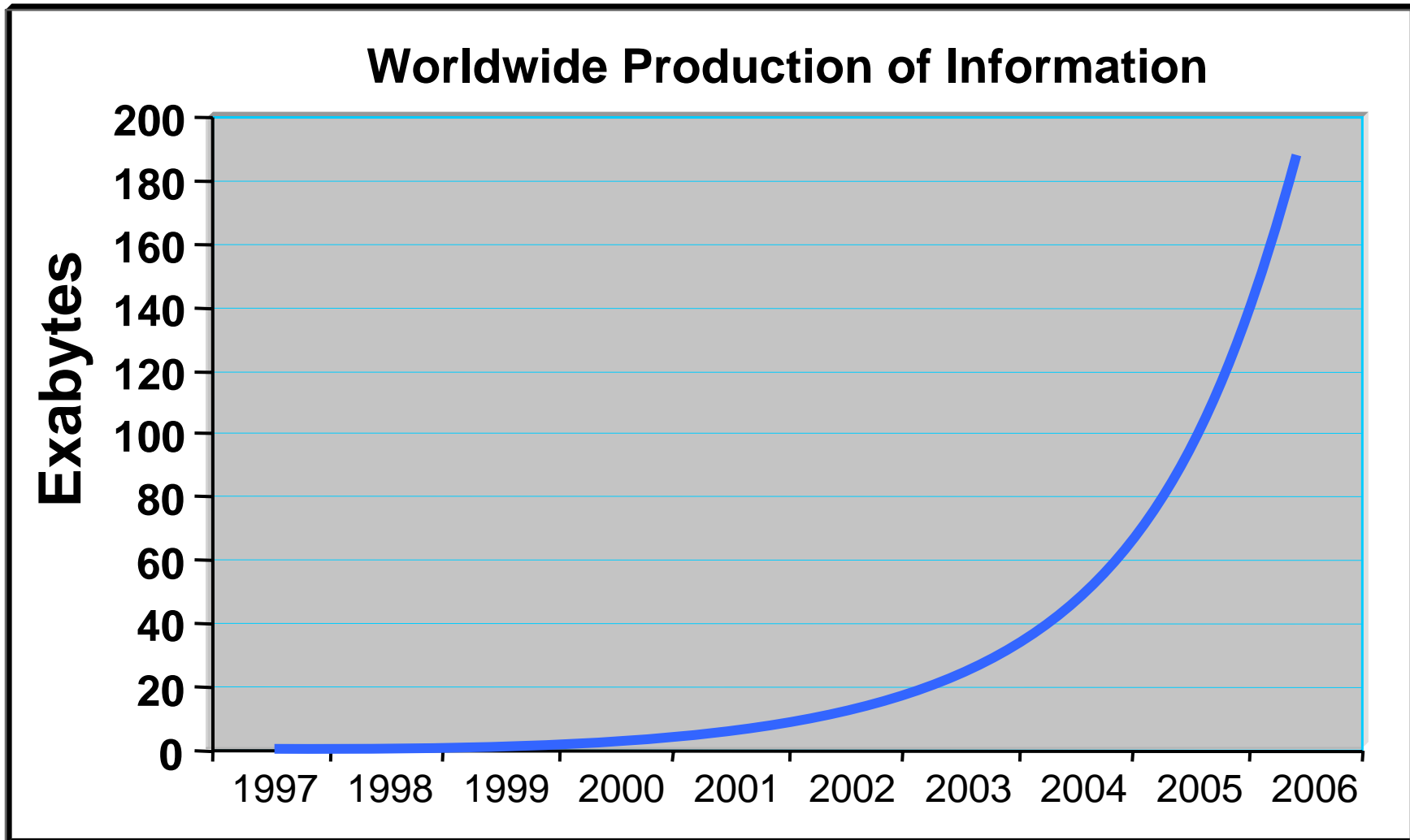
- Technologies supporting e-Science

**Crispin Keable**

**HPC Technical Architect**

**SGI UK**

# The Data Explosion



Source: Gartner Group

**SGI's approach to the Grid is completely formed by our work with "big" data. We believe:**

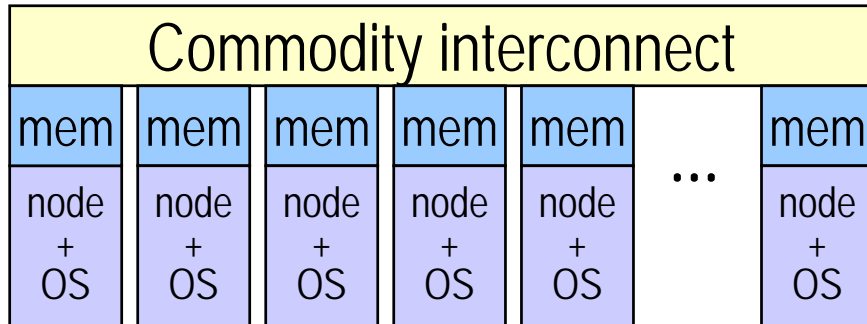
- ***that data is fast becoming too big to move***
- ***that access should not be constrained by the location of the user, the data, or the functionality required to deal with the data***
- ***that the model should be to move the functionality to the data, and provide ubiquitous read and write access to the user with full visualization of results where required***

- **Scaleable Linux Systems**
  - The ability to Share memory across large superclusters
- **Visual Area Networking**
  - The ability to do remote collaborative visualization from high end graphics servers
- **Wide Area File sharing**
  - The ability to share large file systems over very wide areas (up to 8,000 KM)

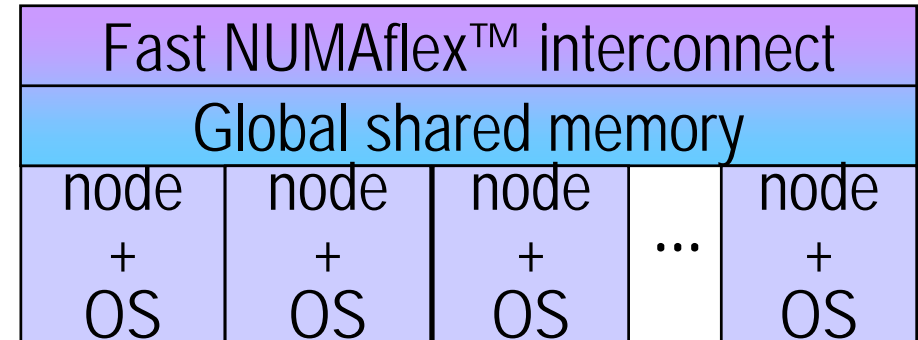
# The Benefits of Shared Memory



## Traditional Clusters



## SGI® Altix™ 3000



## What is shared memory?

- All nodes operate on one large shared memory space, instead of each node having its own small memory space

## Shared memory is high-performance

- All nodes can access one large memory space efficiently, so complex communication and data passing between nodes aren't needed
- Big data sets fit entirely in memory; less disk I/O is needed

## Shared memory is cost-effective and easy to deploy

- The SGI Altix 3000 family supports all major parallel programming models
- It requires less memory per node, because large problems can be solved in big shared memory
- Simpler programming means lower tuning and maintenance costs

# Visual Area Networking

- VAN accelerates that workflow by eliminating travel or data copying while increasing the capabilities of individuals and teams



# SGI CXFS Grid File system demonstration



HPC Linux Server



Sun Server



AIX Server



LightSand S-600  
FC/SONET Gateway

Adtech AX/4000  
WAN Simulator



LightSand S-600  
FC/SONET Gateway

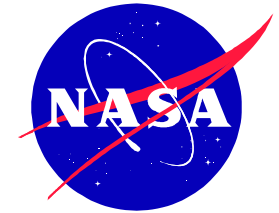


SGI SANserver 1000  
with CXFS

- 100 Mb Ethernet (Metadata)
- 1Gb Fiber Channel (data)
- 622 Mb OC-12 SONET (Metadata & data)

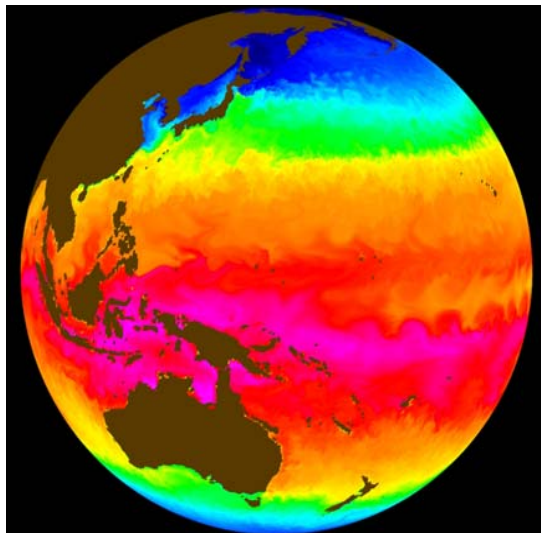
# ECCO Code Performance

## 11/04/03

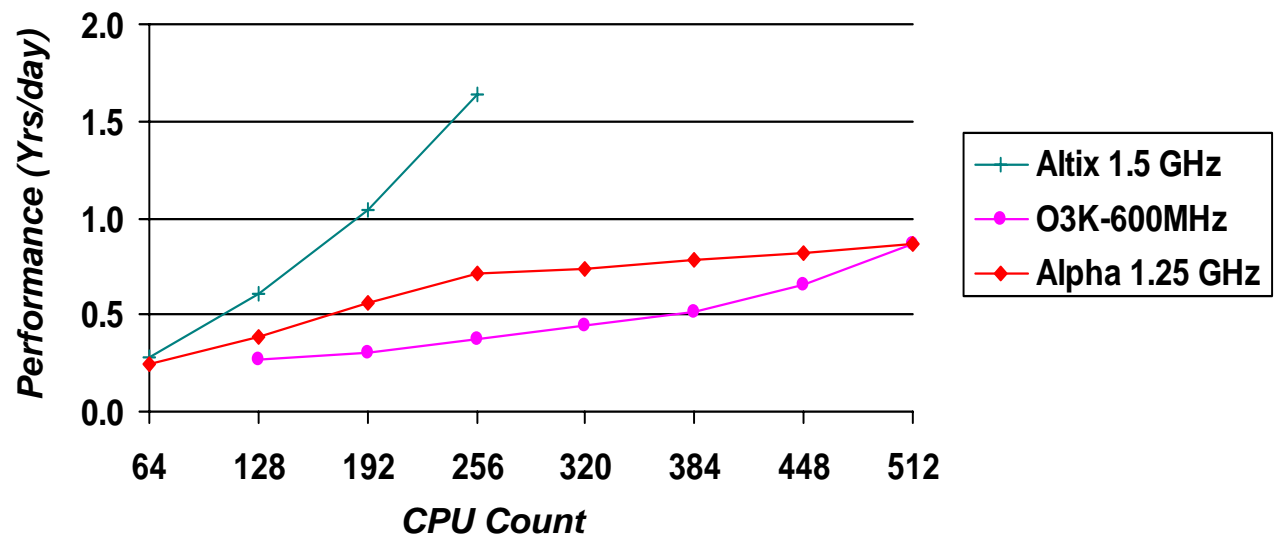


The ECCO code is a well known ocean circulation model, with features that allow it to run in a coupled mode where land, ice, and atmospheric models are run to provide a complete earth system modeling capability. In addition the code can run in a "data assimilation" mode that allows observational data to be used to improve the quality of the various physical sub-models in the calculation. The chart below shows the current performance on the Altix and other platforms for a "1/4 degree" resolution global ocean circulation problem. (in reality, much of the calculation runs at an effective much higher resolution due to grid shrink at the poles).

Note: Virtually no changes to the code have been made across platforms. Only changes needed to make it functional have been done. The preliminary Altix results are very good to date. A number of code modifications have been identified that will significantly improve on this performance number. NOTE: The performance on both Chapman and Altix with full I/O are super-linear. That is as you add more CPUs you get even faster speedups. The Alpha numbers show a knee at 256 CPUs.



**CURRENT PERFORMANCE: 256p Altix 1.5GHz = 1.6 yrs/day !!!! (or a decade a week)**

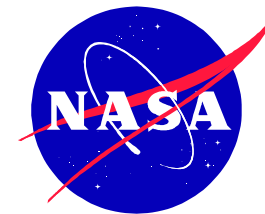


NOTE: Alpha Data re-plotted from Gerhard Theurich charts in NCCS paper



# LEVLER Routine - Results of Dynamic Balance

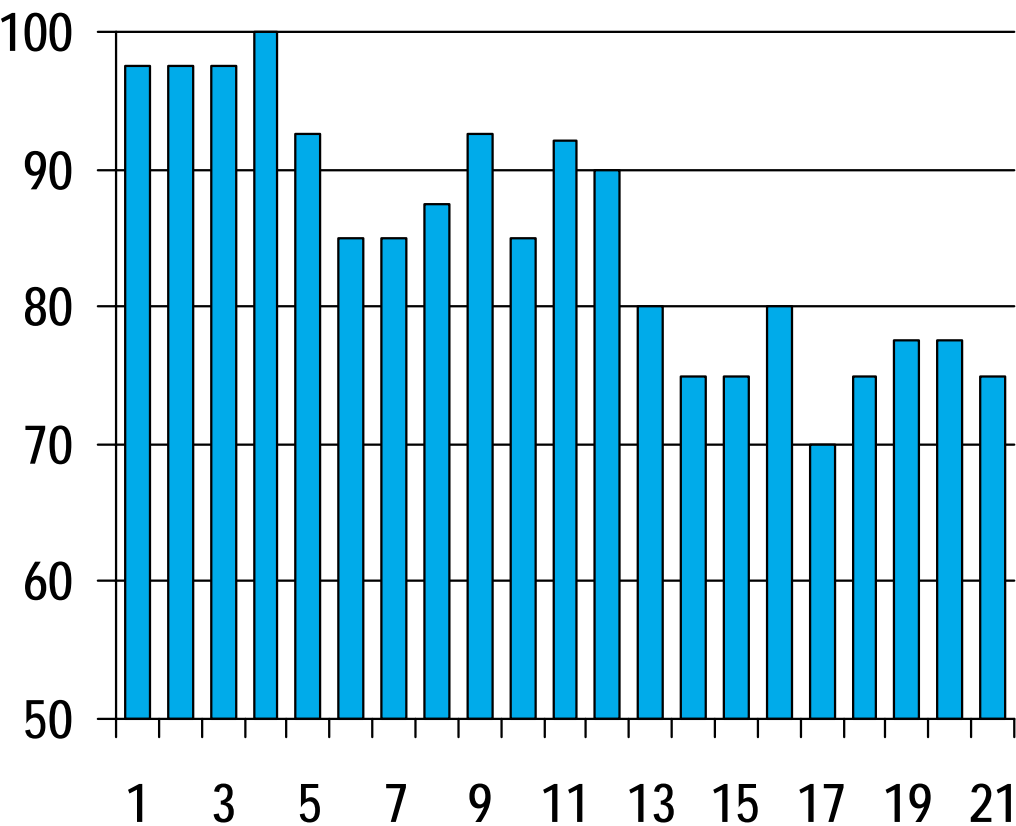
Normalized Wallclock times versus MLP Process Number



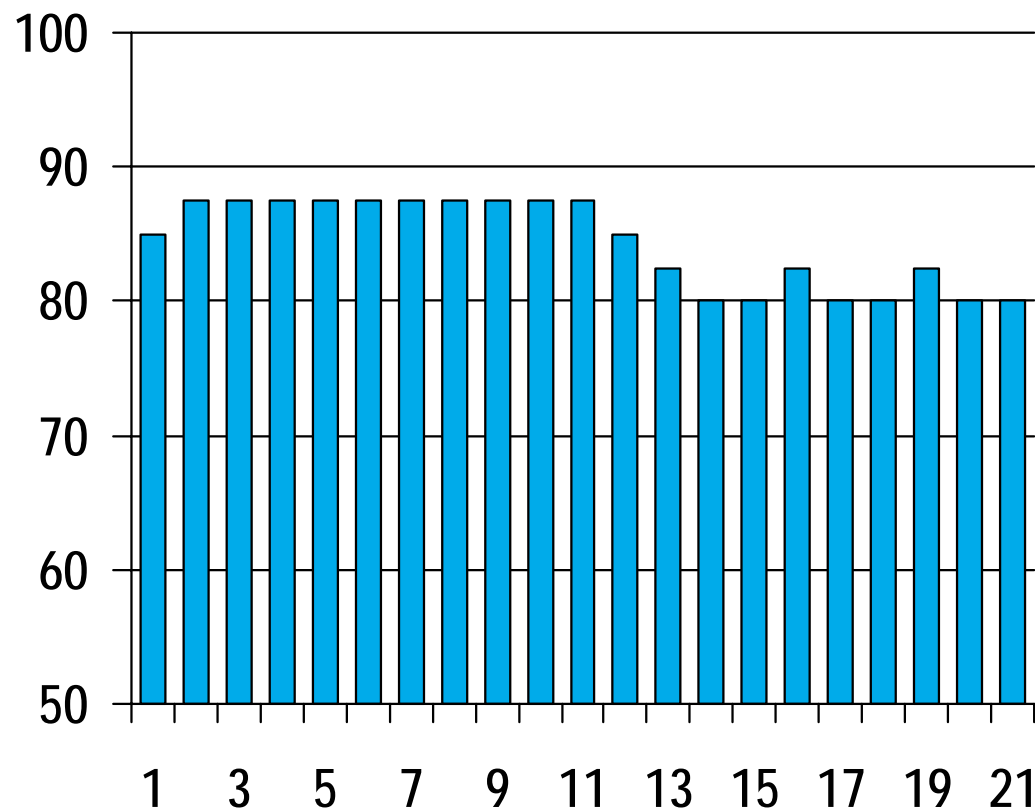
Ames Research Center

Levler off: Wallclock time=720

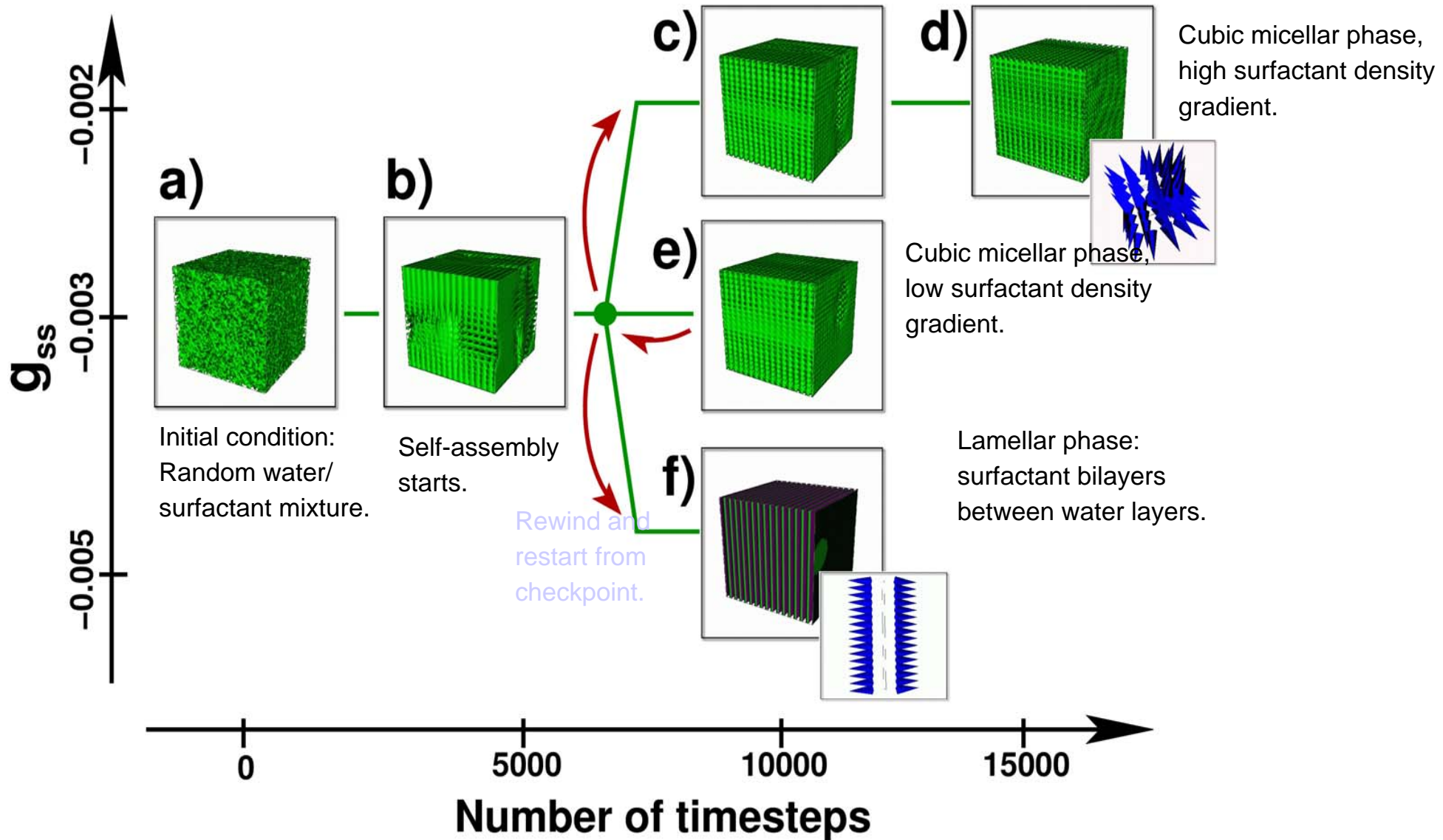
Levler on: Wallclock time=642



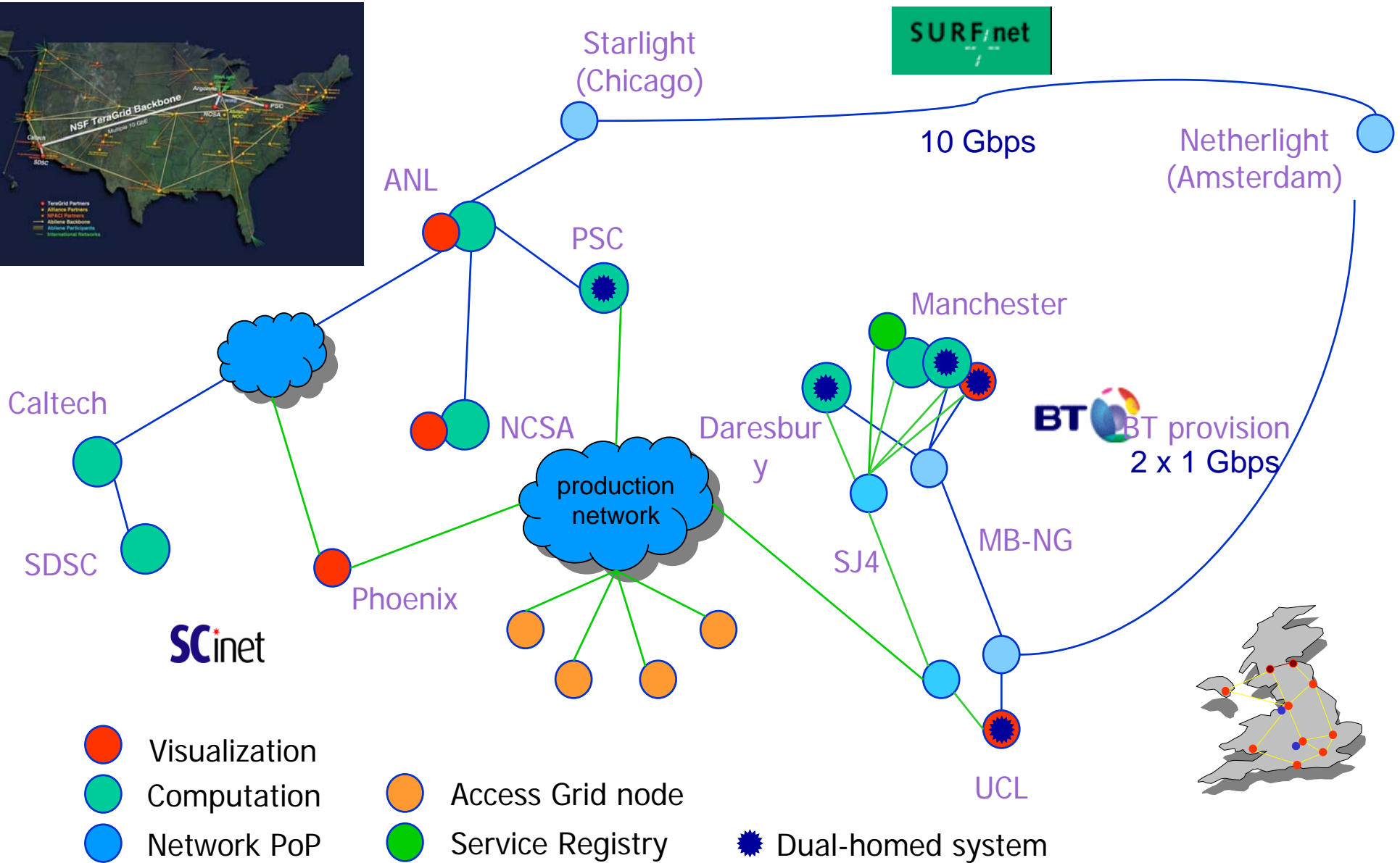
MLP Process Number



MLP Process Number



# TeraGyroid Grid



# Accelerating Workflows: Data Growth Creates Problems

## Manufacturing Example

### Today

1M cells

7 variables

8Bytes/variable

1,000 time steps

Total = 56GBytes

### In 2 to 3 Years

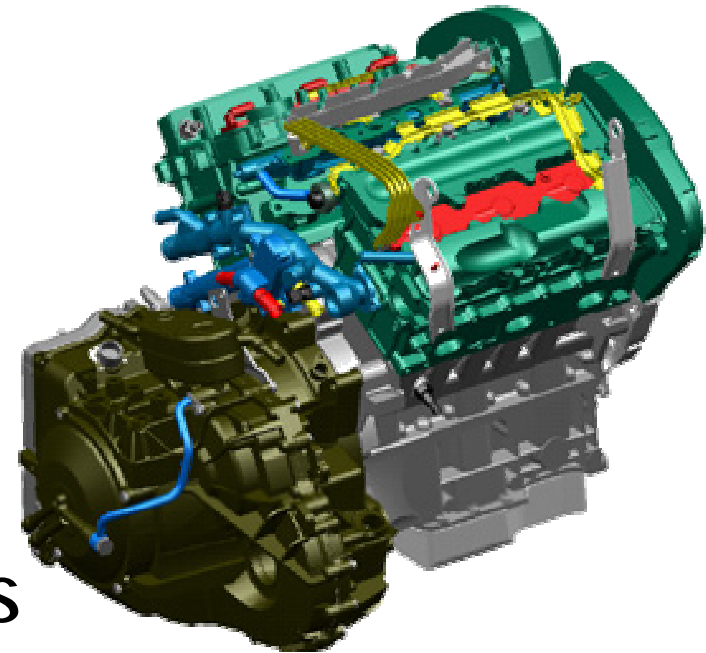
5M cells

12 variables

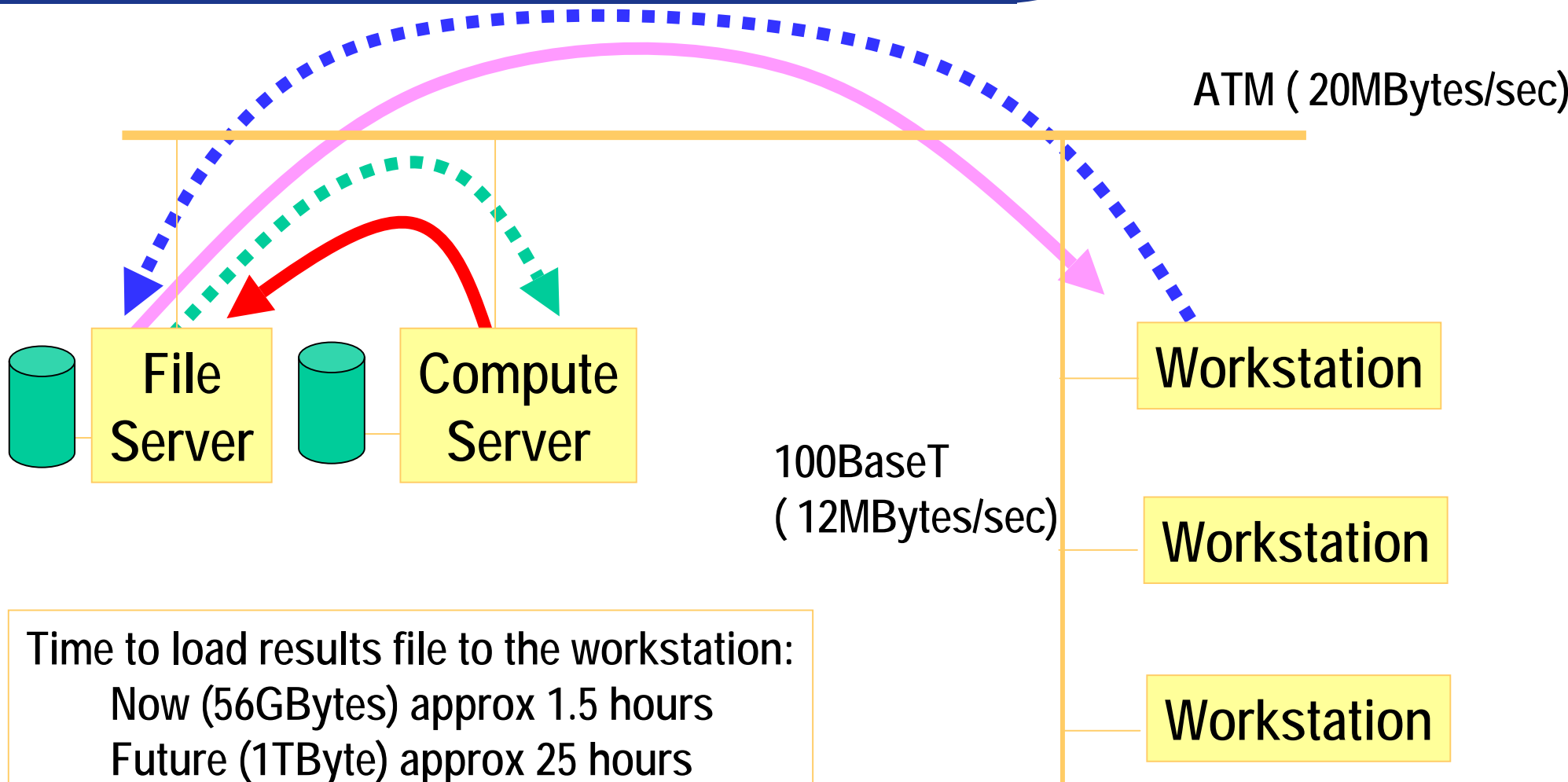
8Bytes/variable

2,000 time steps

Total = 1TBytes



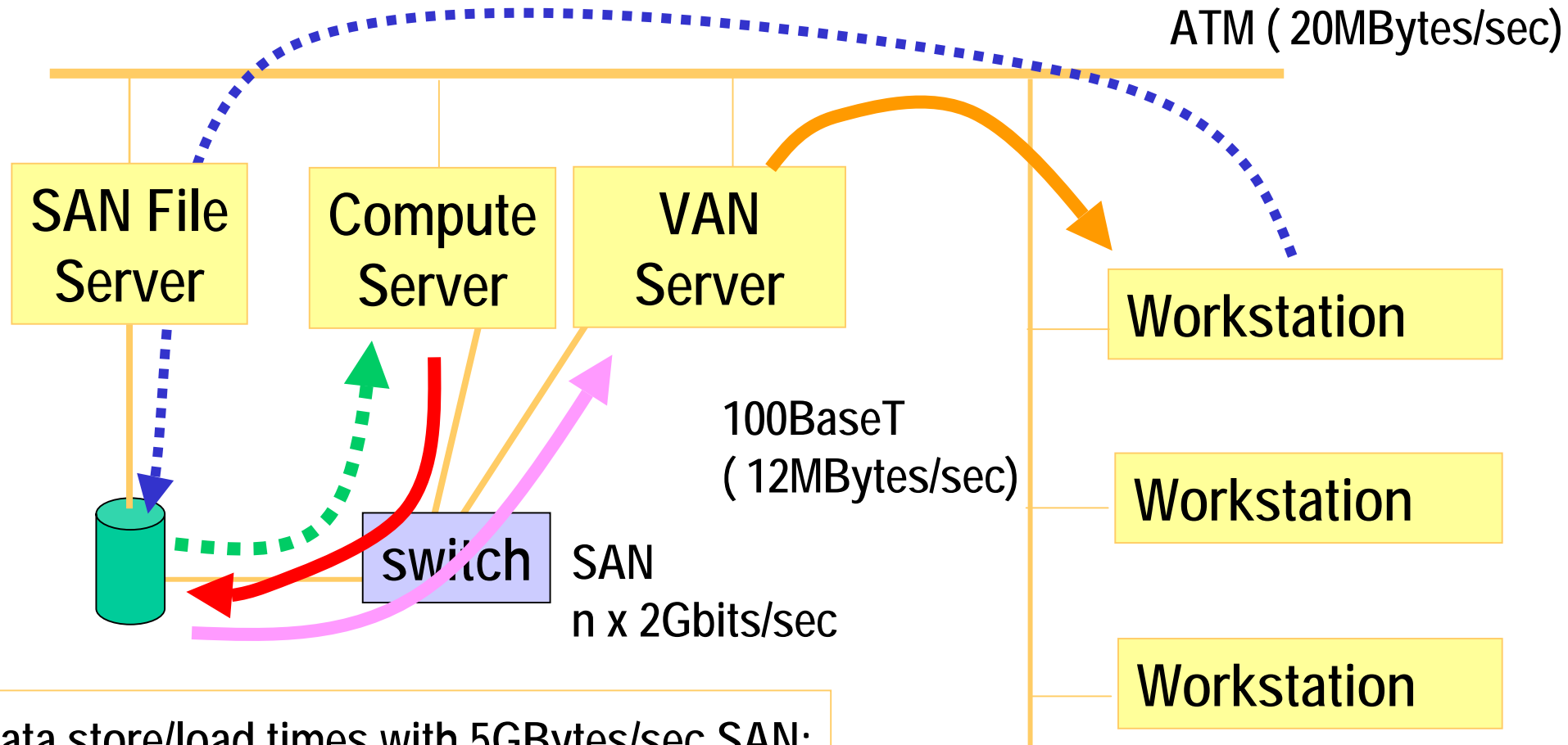
# Accelerating Workflows: Dataflow Problems without VAN



Time to load results file to the workstation:  
Now (56GBytes) approx 1.5 hours  
Future (1TByte) approx 25 hours  
Total transfer times (2 transfers):  
Now (112GBytes) approx 3 hours  
Future (2TBytes) approx 41 hours

**Problem:**  
Data copies slow analysis

# Accelerating Workflows: Dataflow Solutions with VAN



Data store/load times with 5GBytes/sec SAN:  
Now (112GBytes) 11 seconds = 450x faster  
Future (2TBytes) 2.5 minutes = 730x faster  
Spend more time thinking, less waiting !

**Solution :**

**VAN allows Onyx<sup>®</sup> family systems to be shared and distributed**

- The grid should be built not invented
- The rate of data growth is out-pacing the rate of network growth
  - networks are getting slower!
- SGI provides unique technologies to build functional grids today
  - shared memory HPC
  - Visual Area Networking
  - Storage management

sg*i*<sup>®</sup>



Igniting Innovation  
and Leadership