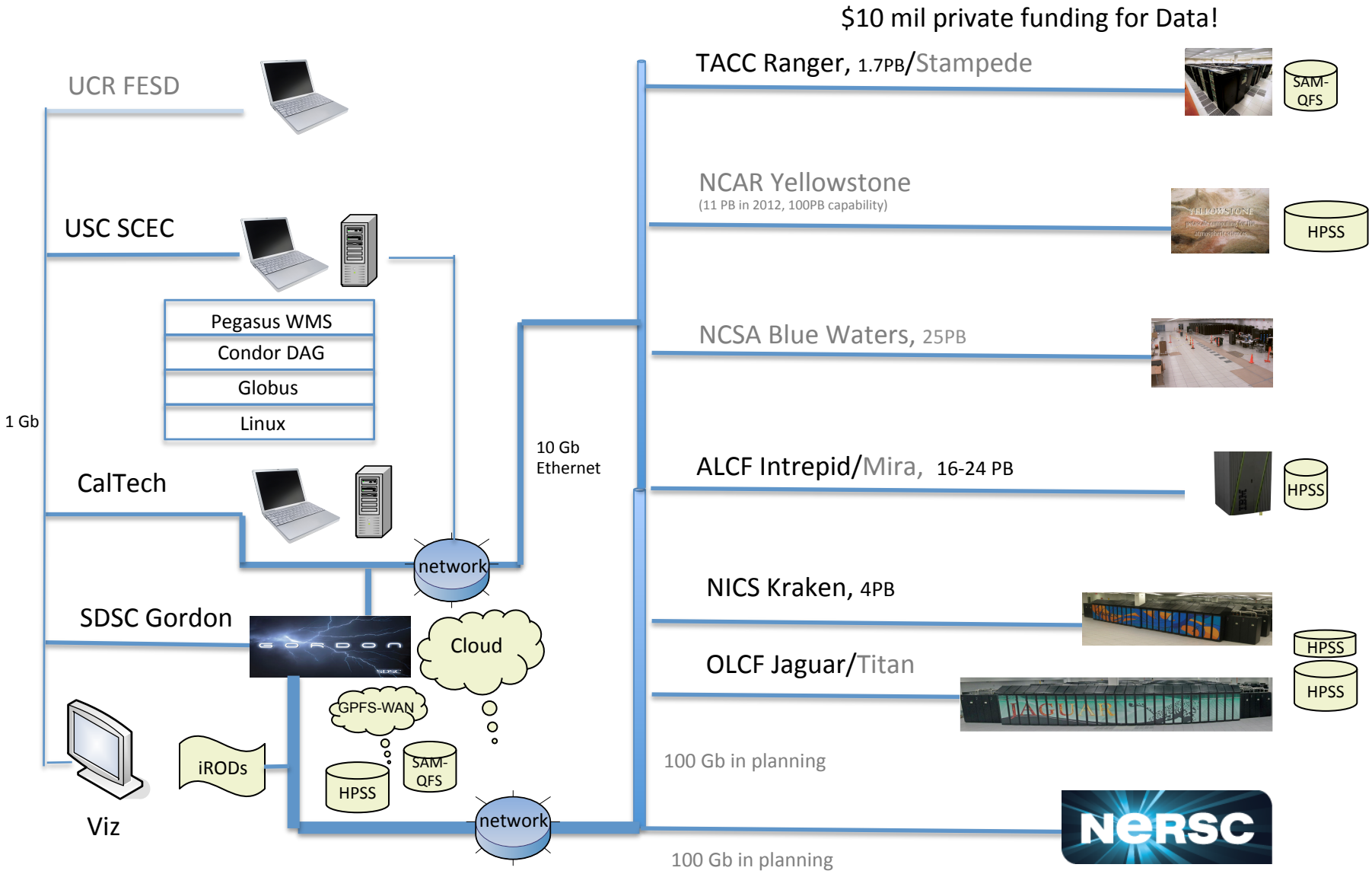
SCEC Large-scale Simulation Data @ SDSC and HPC Centers

Yifeng Cui

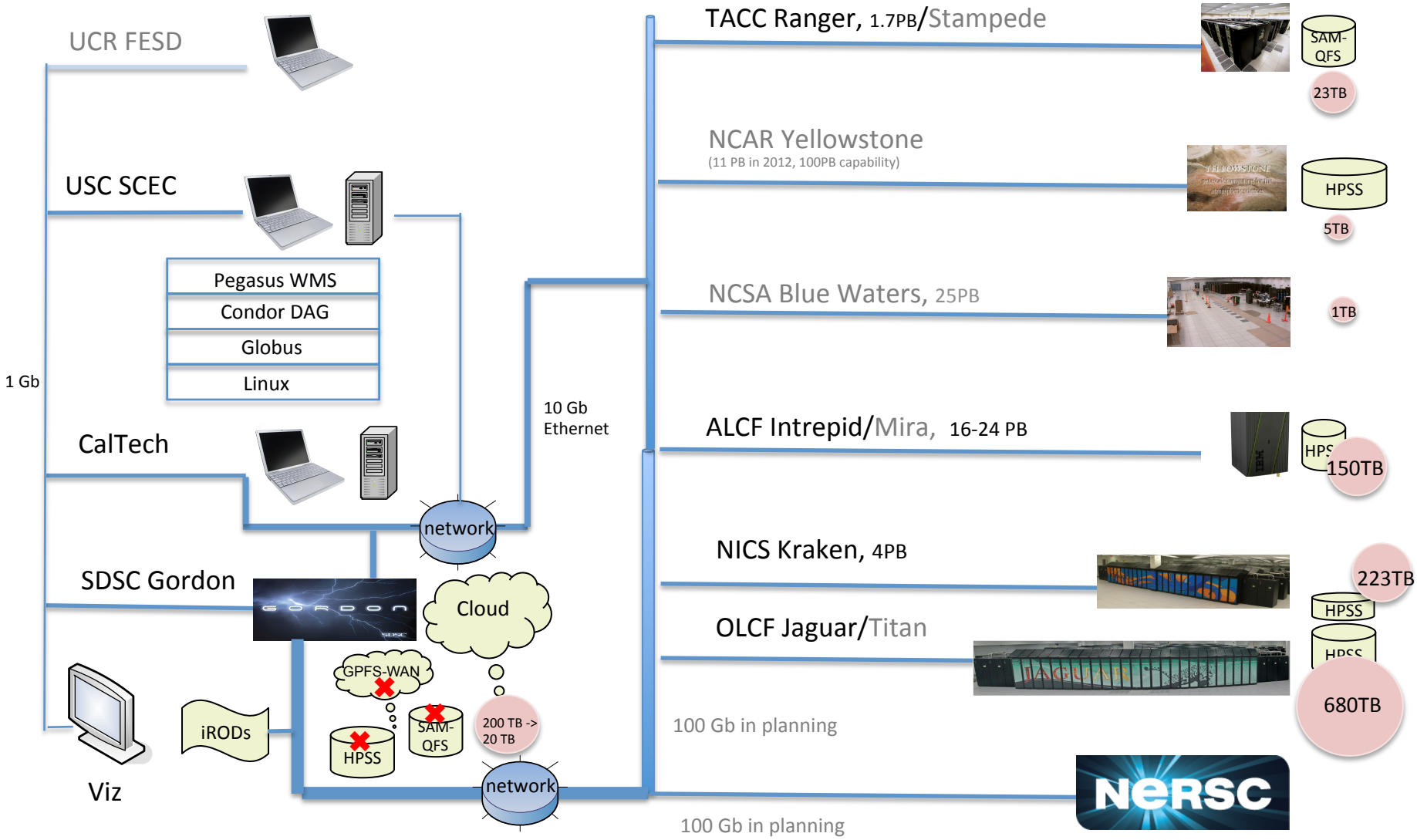
SCEC Simulation Data



SCEC Simulation Data

Parallel IO, Data parking, Data processing, cross-site transfer

$10 mil private funding for Data!

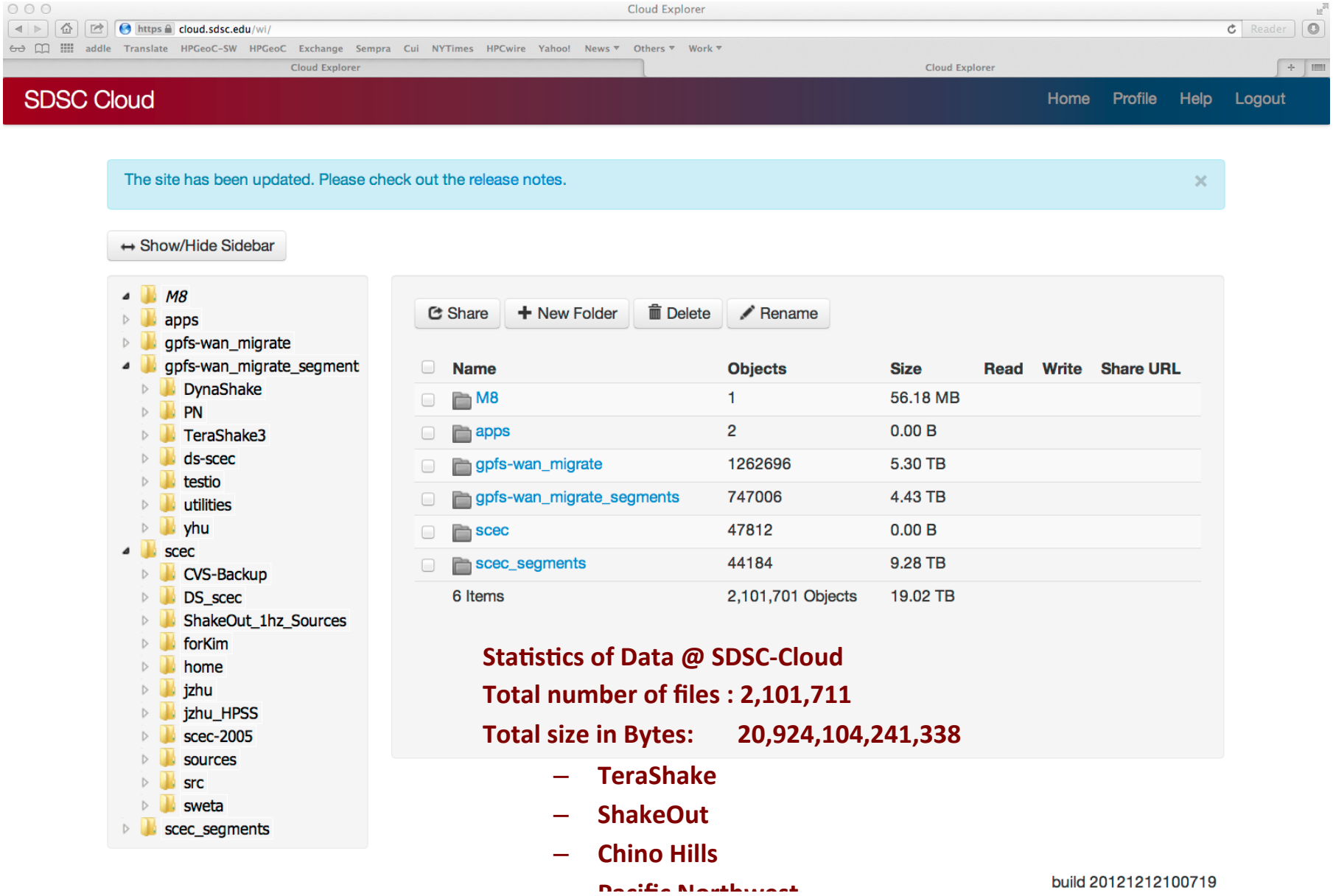


SCEC Visualization Data

SIMULATION		IMAGES			MOVIES		
Name	Count	Count	Size (GB)	Average (KB)	Count	Size (GB)	Average (MB)
TeraShake	7	302560	69.39	34.36	271	13.25	7.15
ShakeOut	4	143315	101.70	186.03	109	18.76	44.06
M8	3	106679	104.83	343.47	68	8.91	44.72
TOTAL	**14**	**552554**	**275.92**	**563.85**	**448**	**40.92**	**95.93**
Website	**1014**	**1014**	**0.66**		**505**	**34.64**	

(Source: Chourasia, SDSC)

SCEC Data @ SDSC Cloud



SDSC Cloud Data Migration

- SDSC starts charging SCEC Cloud storage April 1st, at cost of $3.25 a month per 100GB of storage → $7,800 per year
- SDSC Cloud is not really HPC … plan to delete everything from cloud by end of March
- We purchased 12 TB external hard drives, and move data to the hard drive temporally

  1. running the 'swift.py' (python script to access the openstack swift storage data) to download the scec data to some linux host disk space eg. srbbrick12.ucsd.edu:/data1/sheauc/

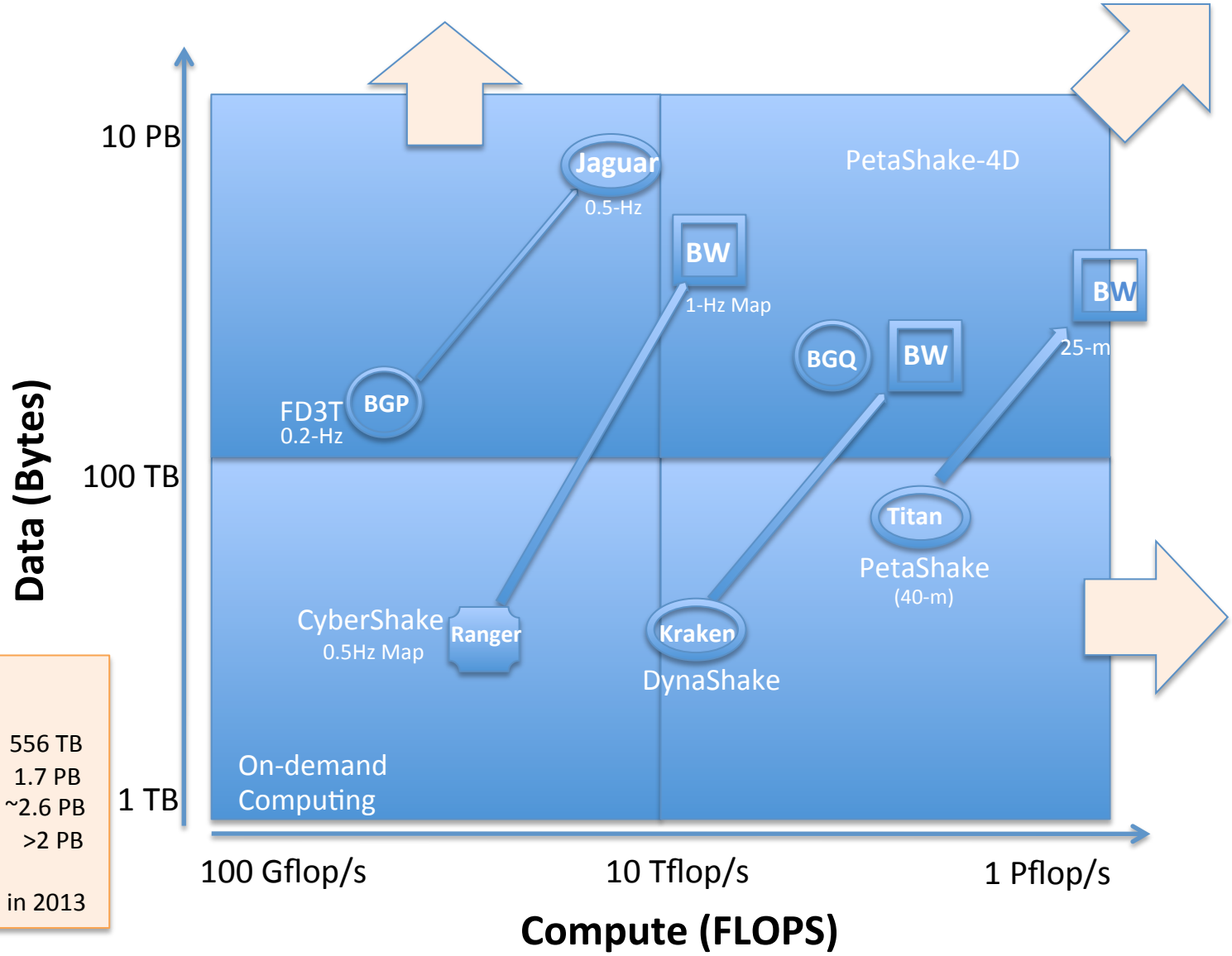
  2. External hard drives are reformatted on Mac OS 10.6. Will use these external hard drives, mounted on Mac, then using 'scp' to copy downloaded cloud data (eg. Srbbrick12) to the external hard drive.

- Long term plan is to transfer all data to USC (which server?)

Where we stand now

- Simulation data increasing both in volume and in value
- Data mobility is a pain point
- Standard time evolution of field variables based simulations which write out time histories
- Parallel I/O, extreme I/O, data parking, and post-processing

SCEC Data Needs



2013-2014:

DynaShake: 556 TB
High-F: 1.7 PB
CyberShake: ~2.6 PB
F3DT: >2 PB

>100 mil Sus in 2013

Short-term Data Management Need

- SCEC standards for naming and storage policies are needed for simulation data collections.
- Challenge to manage storage associated with allocations on both XSEDE and INCITE systems.
- Data maintenance and transfer difficult, backup favored only important simulation data, no derived products.
    - Earthquake rupture time series?
    - Earth structure models?
    - Geo-referenced ground motion time series?
    - Simulation settings, source code, etc
- Dataset catalog management software needed for community access, iRODS, or others?

Future Data Need

- Large data storage needed both for permanent and temporary data sets
- Reliable and high data-rate data transfer
- Robust connectivity to large quantities of data
- 10s of gigabits for collaborative visualization and mining of large data sets
- Authenticated data streams for easier site access through firewalls
- Robust networks supporting distributed simulation – adequate bandwidth and latency for remote analysis and visualization of massive datasets
- Quality of service guarantees for distributed simulations
- Bulk transfer of resultant data sets to SDSC for analysis
- Remote steering, remote visualizations

INPUT and OUTPUT FORMATS of WAVE CODES

Heming Xu

INPUTS

Code (Fortran /C)	nvar	Vp	Vs	ρ	Qp	Qs	γ	Number of mesh file(s) and grid points
AWP-ODC (F)	8							1, 8*nx*ny*nz (3 are not used)
	5							1, 5*nx*ny*nz
	3				CALC	CALC		1, 3*nx*ny*nz
SORD (F)	1	x						1, nx*ny*nz
	1		x					1, nx*ny*nz
	1			x				1, nx*ny*nz
	1						x	1, nx*ny*nz
SPEM (F)	4							1)Homogenous block material model 2)Extra variables: material_ID, anistropic_flag,domain_ID
AWP-Graves (C)								Layer structure

INPUTS

	nvar	Vp	Vs	ρ	Qp	Qs	γ	Number of mesh file(s) and grid points
ShuoMa (F)	3							Left/right to fault (uniform)

nvar: # of input variables; CALC: calculated variables from others
Vp: P wave speed; Vs: S wave speed; ρ: density
Qp, Qs: attenuation for P and S
γ: Kevin-Voigt viscosity parameter
nx,ny,nz: grid dimension

OUTPUTS:

	Vx	Vy	Vz	
AWP-ODC	Individual file	Individual file	Individual file	Seismogram/ movie
SORD	Individual file	Individual file	Individual file	Seismogram/ movie
SPEM3D	Individual	Individual file	Individual file	seismogram
				movie
AWP-Graves				Seismogram/ time slice
ShuoMa	Individual file	Individual file	Individual file	seismogram

1) To minimize the code change and to have flexibility, the input variables in different files seem to a better choice.

2) Outputs always are done in different files in AWP-ODC and SORD for different variables. The formats need to be standardized.