Regression and Time Series Analysis of the SCD Mass Store System

Douglas Nychka,

www.image.ucar.edu/˜nychka

- Data and Informal Conclusions
- Relationship of Storage to FLOPS
- Relationship of Read access to storage and FLOPS
- Extrapolation of storage based on past history

Supported by the National Science Foundation

MSS task force 11/2005

Provenance

Detailed data supplied by Tom Engle and DASG

Dataset: factors and levels

- approximately 20 largest “groups” *(NCAR, WACCM, VETS etc)*
  Key analysis group formed as sum of NCAR, CSL and UNIV and termed USERS
- time *monthly data from 1999-2005*
- activity *(storage, create, read, write)*
- measures *(Gb, gaus, access)*

Also monthly *Sustained GFlops* and *Unique total MSS size* were supplied by Gene H.

Patterns in activity that would useful for MSS planning

Based on a series of meetings the key features are

- *Storage as a function of compute capacity.*
  This was simplified to be the dependence of the USER storage on the sustained GFLOPS available through SCD.
- *Read access as a function of archive size*
  This simplified to be the number of monthly read accesses for the USER group asa function of storage.

There several other aspects of the MSS that effect its function but these two appear to be the most serious constraints.

The strategy:

Forecast monthly storage rates based on planned compute capacity (in GFLOPS)

Infer the read accesses from the forecasted archive growth and GFLOPS . Based on conversations with John M. and Eric T. other attributes of the MSS are not as crucial for planning.

Some conclusions

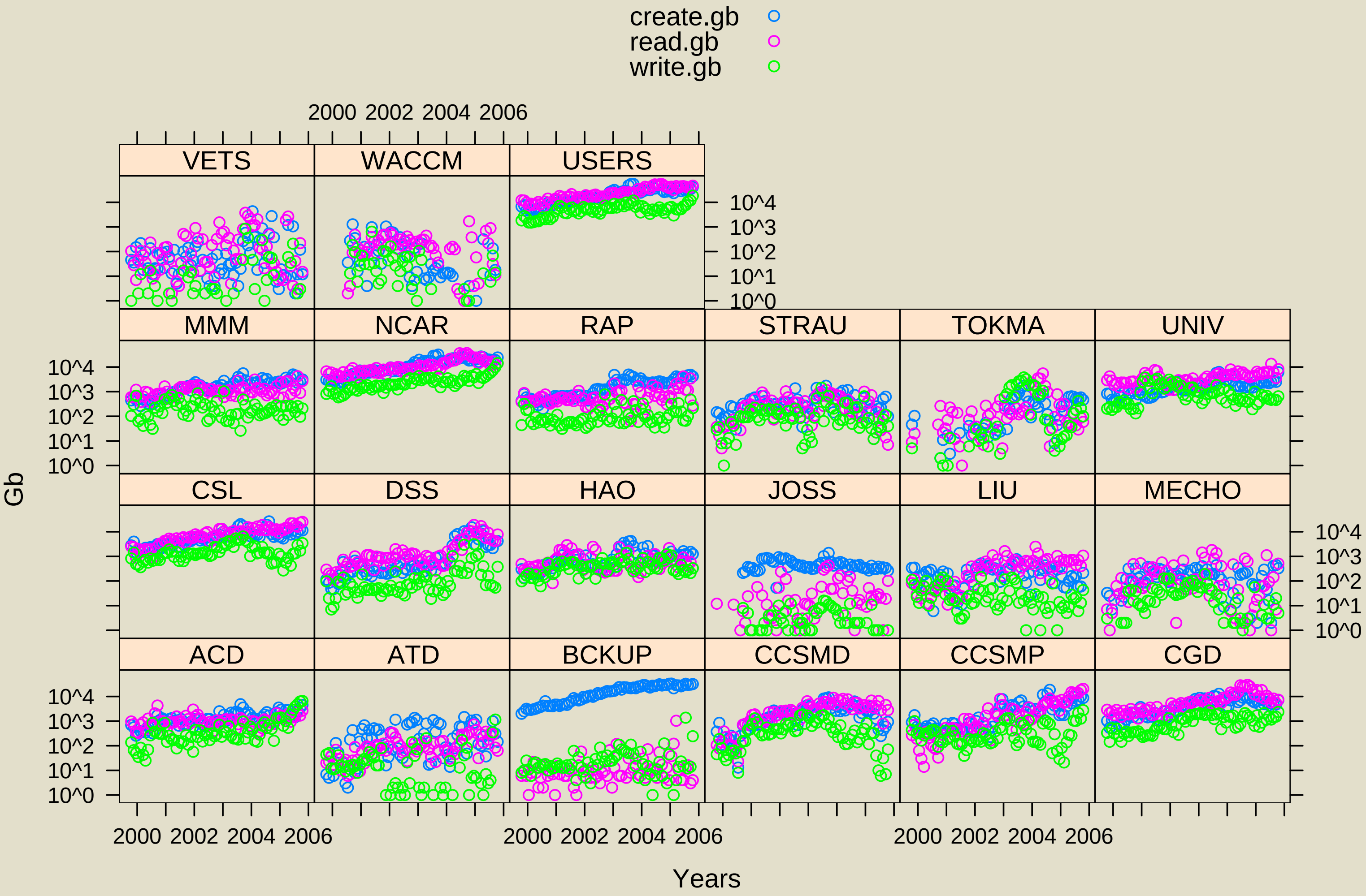
- The USER and Total MSS storage can be predicted using sustained FLOPS. e.g. a likely *overestimate*:

$$\textit{Monthly USER storage (Gb)} = 6500 + 28.4(\textit{ GFLOPS})$$

$$\textit{Unique MSS (GB)} = 8325 + 31.7(\textit{GFLOPS})$$

- Monthly storage is decreasing as a function of sustained GFLOPS but also has substantial variability.
- Read accesses for the USER group is decreasing in proportion to the archive size ($\leq$ 400 reads/Pb/month.) but is linearly related to GFLOPS
- The regression analysis should be continually updated with new data.

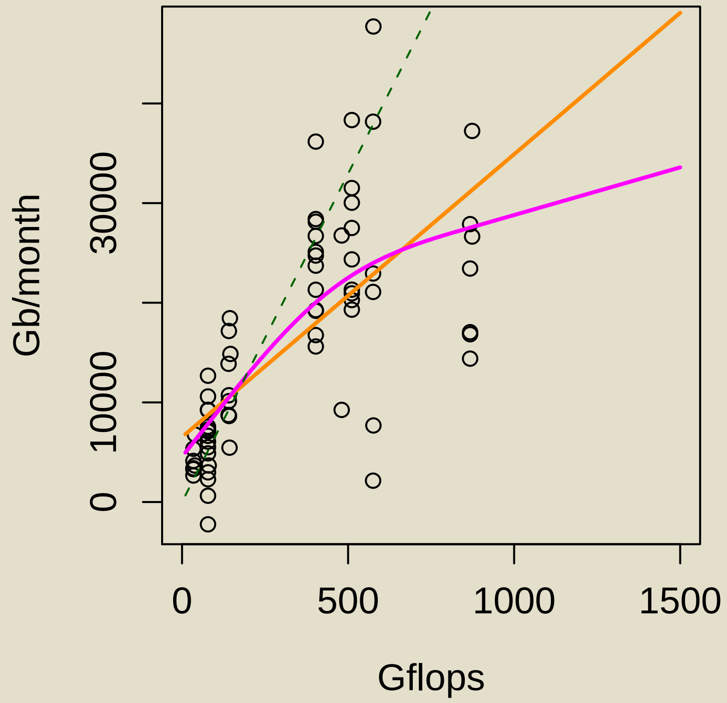
Raw data in Gb over time



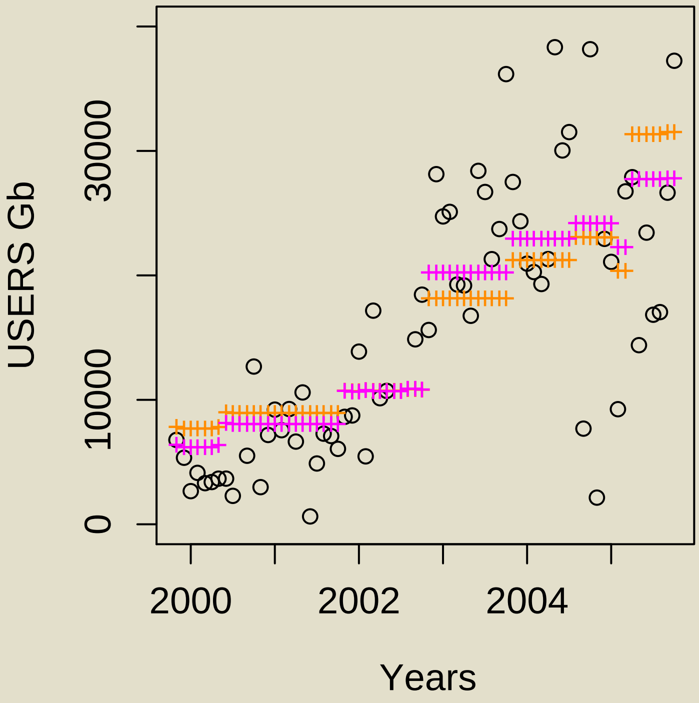
Ginger's favorite plot ...

Monthly Storage to sustained GFLOPs

Regression relation

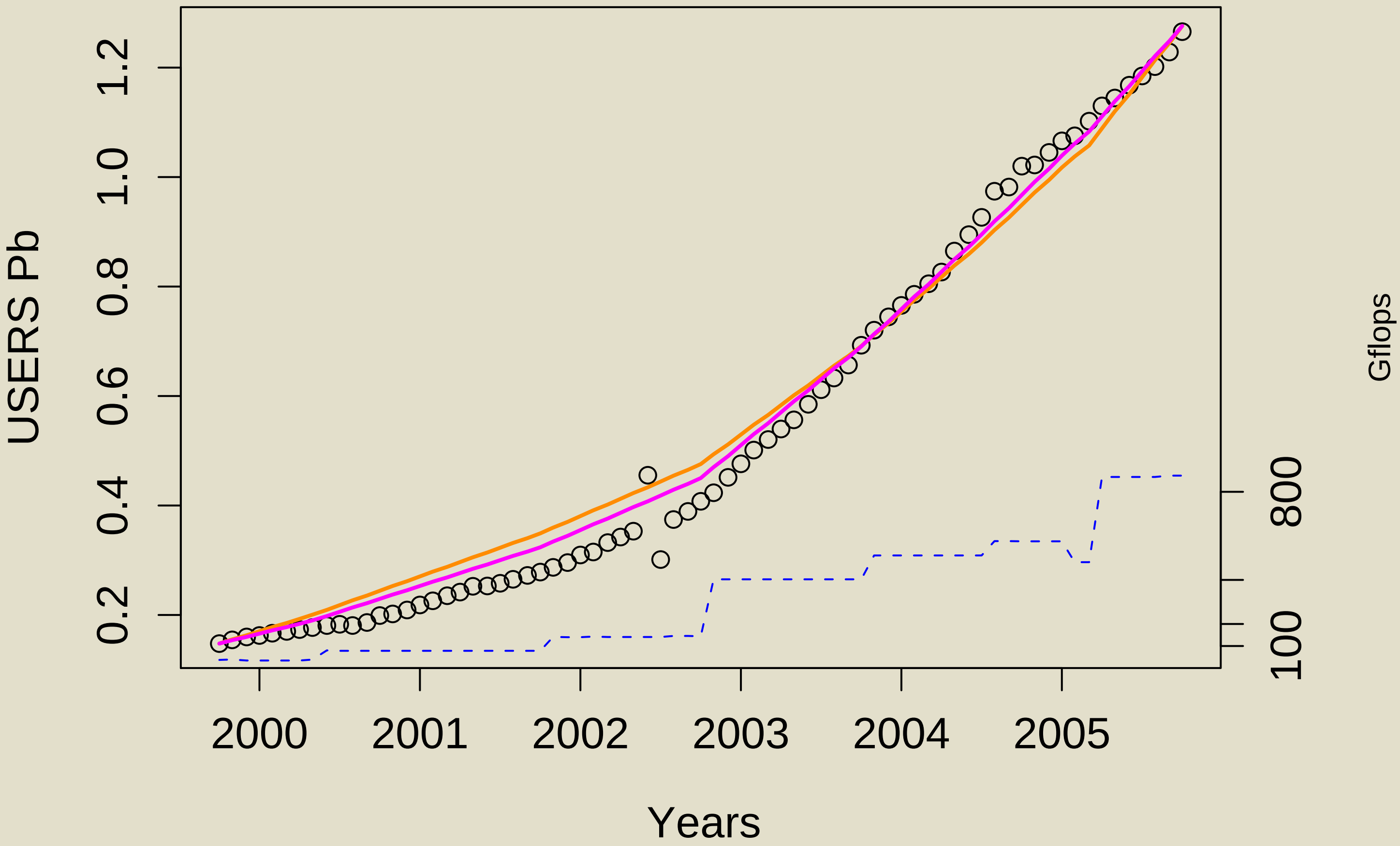


Fitted storage by time



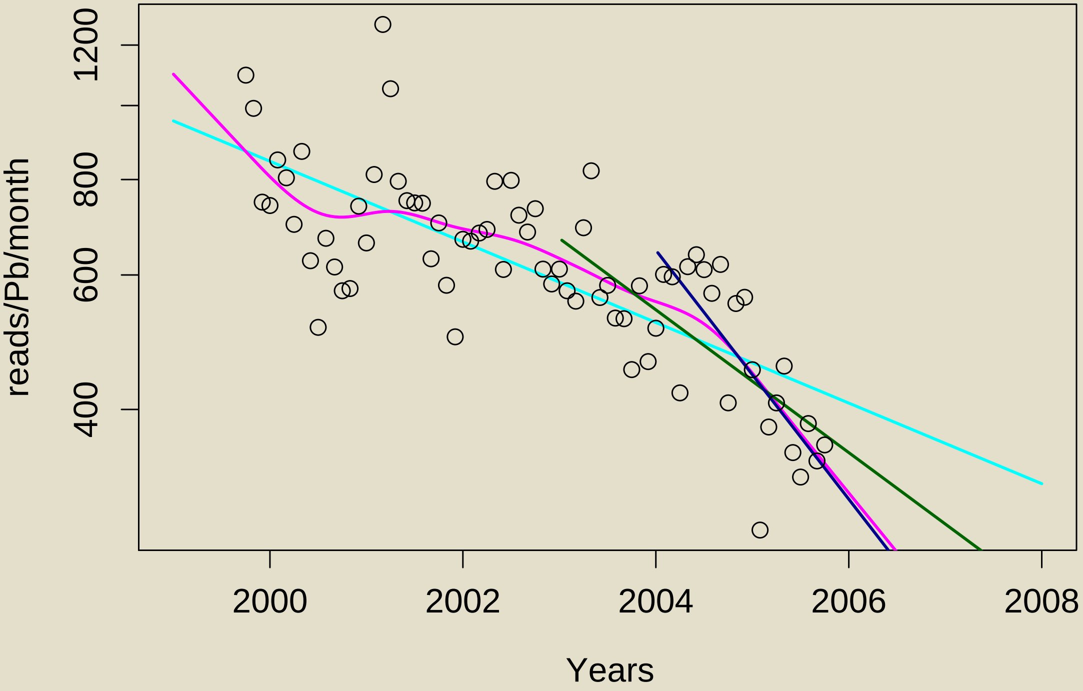
average of bytes/GFLOPS, LS, smoothing spline

USER Storage predicted from GFLOPs



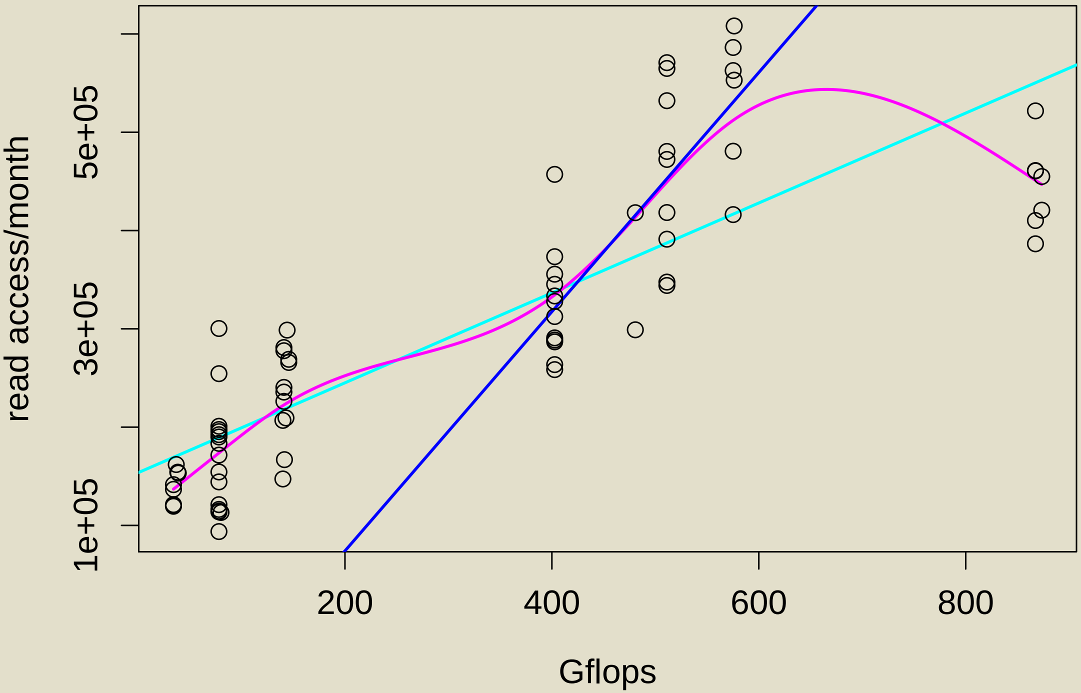
Data, LS, smoothing spline

Read access and archive size



Data, LS, LS > 2003 LS > 2004 smoothing spline

Read access and GFLOPS



Data, smoothing spline

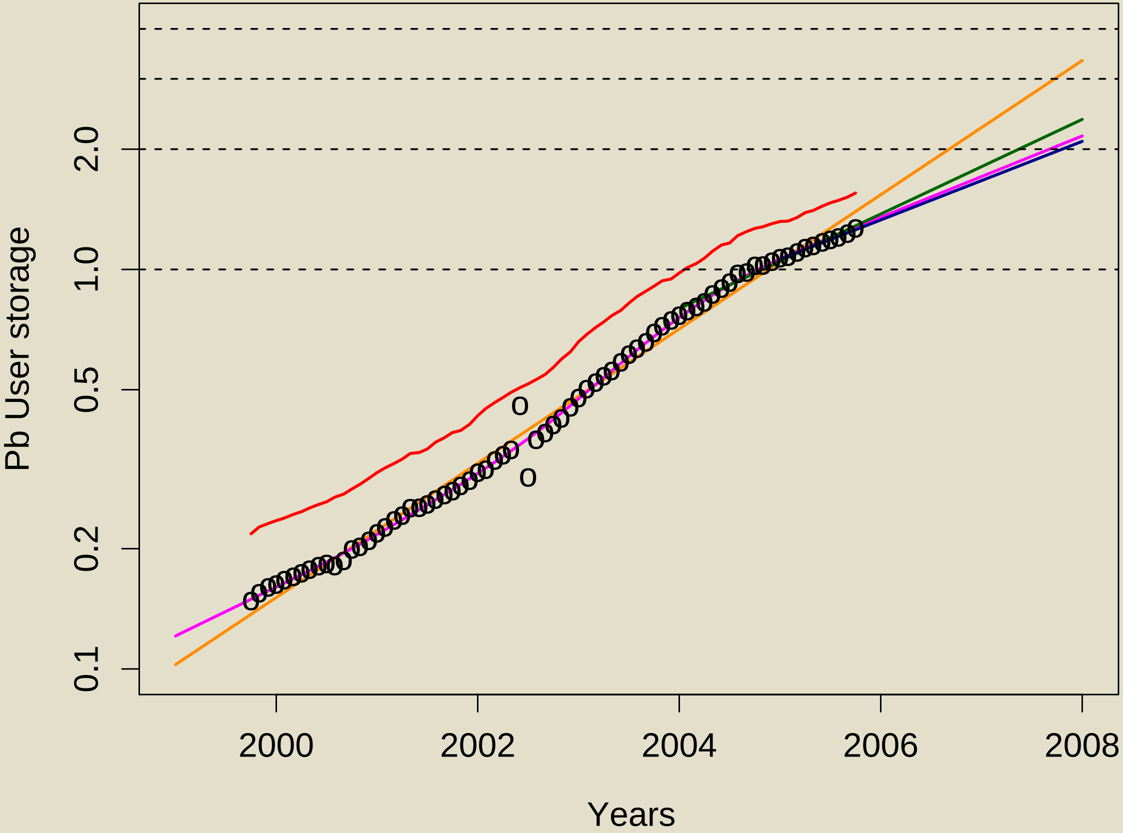
LS: 460 reads/GFLOP

LS for range [300,700]: 1200 reads/GFLOP,

Time forecasts of USER growth (doubling times)

LS (1.75,1.83), LS 24 months (2.4,2.7)

LS 12 months (2.8,3.2)



Total Unique, User, smoothing spline

Thank you!