Grand Challenge of Indiana Water: Estimate of Compute and Data Storage Needs
Indiana University
Beth Plale, 9 April 2016

Summary

This study is undertaken to assess the computational and storage needs for a large-scale research activity to study water in the State of Indiana. It draws its data and compute numbers from the Vortex II Forecast Data study of 2010 [1] carried out by the Data To Insight Center at Indiana University. Detail of the study can be found in each of the archived data products (which contains results of a single weather forecast plus 42 visualizations created for each forecast.) See https://scholarworks.iu.edu/dspace/handle/2022/15153 for example archived data product.

A summary of the compute and storage needs for the project is projected in the table below. The following sections give the methodology used to compute the totals.

Water Grand Challenge; Summary of Compute and Storage Need			
Parallel (model) Computing; yearly need (I/O in Terabytes (TB))			
	SU	41.45 M	Compute hour units per year in millions
	Input	28 TB	Data volume into HPC resource / year
	Output	2,005 TB	Data volume out of HPC resource / year
Indiana Deep Data Well (in Terabytes (TB))			
	yearly growth rate	2,402 TB	growth per year
Analytics HPC Computing (I/O in Terabytes (TB))			
	SU	4.15 M	Compute hour units per year in millions
	Input	28,200 TB	Data volume into analytics HPC resource / year
	Output	2.82 TB	Data volume out of analytics HPC resource / year

I. Methodology

The model used in this study is the WRF weather research forecast model as used in the LEAD II effort of [1]. WRF was run 5-6 times a day to generate a short-term 24-hour forecast. The resolution is about 13 Km resolution, and the forecast was run over a region about the size of a large state. That is, it was sized to run fast. It ran on 1024 cores and the end-to-end time from forecast initiation to delivery of visual products on cell phones was restricted to 50 minutes.

The study uses per forecast numbers from the above study as a baseline model/simulation run. It then adjusts/extends for use modes, # users, and model limitations to capture the anticipated yearly demand of parallel HPC computing and storage resources on campus. It then, in a similar manner, uses the model/simulation workload estimates to estimate the size of the Indiana Deep Data Well and the workload on Analytics HPC resources.

Per forecast numbers: Per forecast/model numbers are the base units used throughout. They represent a single run of the model as described above on the 1024 cores. SU computed as cores/time so roughly 1024 cores for 32 min = 514 SUs.

Single WRF Forecast:
SU: 514
Input: .25 GB
Output: 1 GB

II. Parallel (model) Computing

Use Mode: The volume of use for HPC compute resources is based on the needs of 1-2 modelers as they anticipate their needs. The model run numbers are as above. To accommodate future demand, and possible limitations of the reference model, we use a faculty growth factor of 12 and a model expansion factor of 4.

1. Continuous: 5 runs/day, single runs
2. Ensembles: 5 runs per month @ ensemble width of 20; runs uniformly distributed over the month. Assume 1 input feeds all runs within a single ensemble run. Assume each ensemble run has its own output.
3. Researcher: 5 runs/day over one 7 day period per month. This is the burst scenario where researcher preparing for a paper submission. Non-uniform distribution, but that our study does not capture daily use so does not use distribution information.

Growth factor: we estimated a growth factor of 12 for faculty growth

Expansion factor: we include an expansion factor of 4 for longer running, larger area, or finer resolution models because our forecasts were short duration.

The base numbers for modes of use are below. These are the numbers before the growth and expansion factors are applied, so all are multiplied by 12*4 = 48.

Continuous:

SU:	79,670 SU/mo
Input:	38.75 GB/mo
Output:	155 GB/mo

Ensemble Runs:

SU:	51,400 SU/mo
Input:	1.25 GB/mo
Output:	100GB/mo

Researcher activity (bursty around paper deadlines):

SU:	17,900 SU/wk & mo
Input:	8.75 GB/wk & mo
Output:	35 GB/wk & mo

Totals: With growth factors applied above activity by mode, resulting use of computational resources is:

Parallel (model) Computing; yearly need (I/O in Terabytes (TB))			
	SU	41.45 M	Compute hour units per year in millions
	Input	28 TB	Data volume into HPC resource / year
	Output	2,005 TB	Data volume out of HPC resource / year

III. Indiana Deep Data Well

The Indiana Deep Data Well is a unique go to resource for water resource data in the state of Indiana. It facilitiates research, education, and outreach. Where data are needed but well preserved elsewhere (DNR, Army Corps of Engineers) the data are linked through a deep well ontology. Where the data are at risk, they are stored to the deep well. Research products of IU are stored as well, with varying levels of quality control and preservation as fits the need.

The size of the Indiana Deep Data Well is expressed as a growth rate by year, with growth estimated by summing the following three kinds of data:

Model Runs: we estimate retention of model runs at 50% of total number of runs. Since a model generated output (1GB) is about the size of the code base around the model, the estimate of saving 50% of model runs could be interpreted as saving 25% of model runs along with a copy of the model. This is 2,004 TB * .50 = 1,002 TB.

Indiana-relevant data at risk: we use the model input data (not output) because it is of a more realistic size for other kinds of data, and multiply that by a factor of 10 to capture heterogeneity expected. This is 28 TB * 10 = 280 TB

Metadata, indexes, etc: we use the volume of data computed as Indiana-relevant data at risk and multiply it by a factor of 4 to capture the storage need for indexing, linking, indexes, ontological info, etc. This is 280TB * 4 = 1,120 TB

Indiana Deep Data Well growth rate = 1,002 TB + 280 TB + 1,120 TB = 2,402 TB

Indiana Deep Data Well (in Terabytes (TB))			
	yearly growth rate	2,402 TB	growth per year

IV. Analytics HPC Workload (e.g., Hadoop or Spark)

An analysis workload is a workload that employs data mining, machine learning, statistical analysis, or text mining on data in the Indiana Deep Data Well. The compute workload is estimated as 10% of the model need for SUs because data analytics is not as computationally intense as is model and simulation activity. The input and output are expected to be at the same magnitude of the model behavior, except that we swap input and output sizes because unlike models, data analytics starts with a large dataset and ends with a small one. Models are usually the opposite.

Analytics HPC Computing (I/O in Terabytes (TB))			
	SU	4.15 M	Compute hour units per year in millions
	Input	28,200 TB	Data volume into analytics HPC resource / year
	Output	2.82 TB	Data volume out of analytics HPC resource / year

References

1. B. Plale, K. Brewster, C. Mattocks, A. Bhangale, E. C. Withana, C. Herath, F. Terkhorn, and K. Chandrasekar. Dataset: Weather Forecast Data from the D2I-Vortex2 project. May 1 to Jun 15, 2010. Bloomington, Indiana: Data to Insight Center. See https://scholarworks.iu.edu/dspace/handle/2022/15153 for example archived data product