Compression By Default – Reducing Total Cost of Ownership of Storage Systems

Michael Kuhn, Konstantinos Chasapis, Manuel F. Dolz, Thomas Ludwig

Scientific Computing
Department of Informatics
University of Hamburg

2014-06-23

- I/O is a major stumbling block on the road to Exascale
  - Storage is an important portion of the TCO
- CPU speed and HDD capacity have increased by factors of 500 and 100 every 10 years, respectively
  - Growth of HDD capacity has started to slow down
  - More investment required to keep up with computation
- Minimize the amount of data to reduce cost
  - HPC storage servers can transparently compress file data
  - Current CPUs provide ample performance for data processing without impacting applications
  - Approaches such as deduplication are too costly

- Provide a model to estimate benefit of compression in advance
  - Only take into account costs of the actual storage hardware
  - Do not consider indirect savings (for example, less cooling and required space) for now
- Analyze the effectiveness for different data sets
  - Climate data can benefit significantly
  - Compression ratios of 1.5 and more are possible
- CPU and power consumption overhead are important
  - Modern algorithms (`lz4`) produce negligible overhead (~1 %)
  - Other algorithms (`gzip`) provide higher compression ratios but significantly increase power consumption
- Store the same amount of data on less storage hardware
  - Consequently, TCO can be reduced

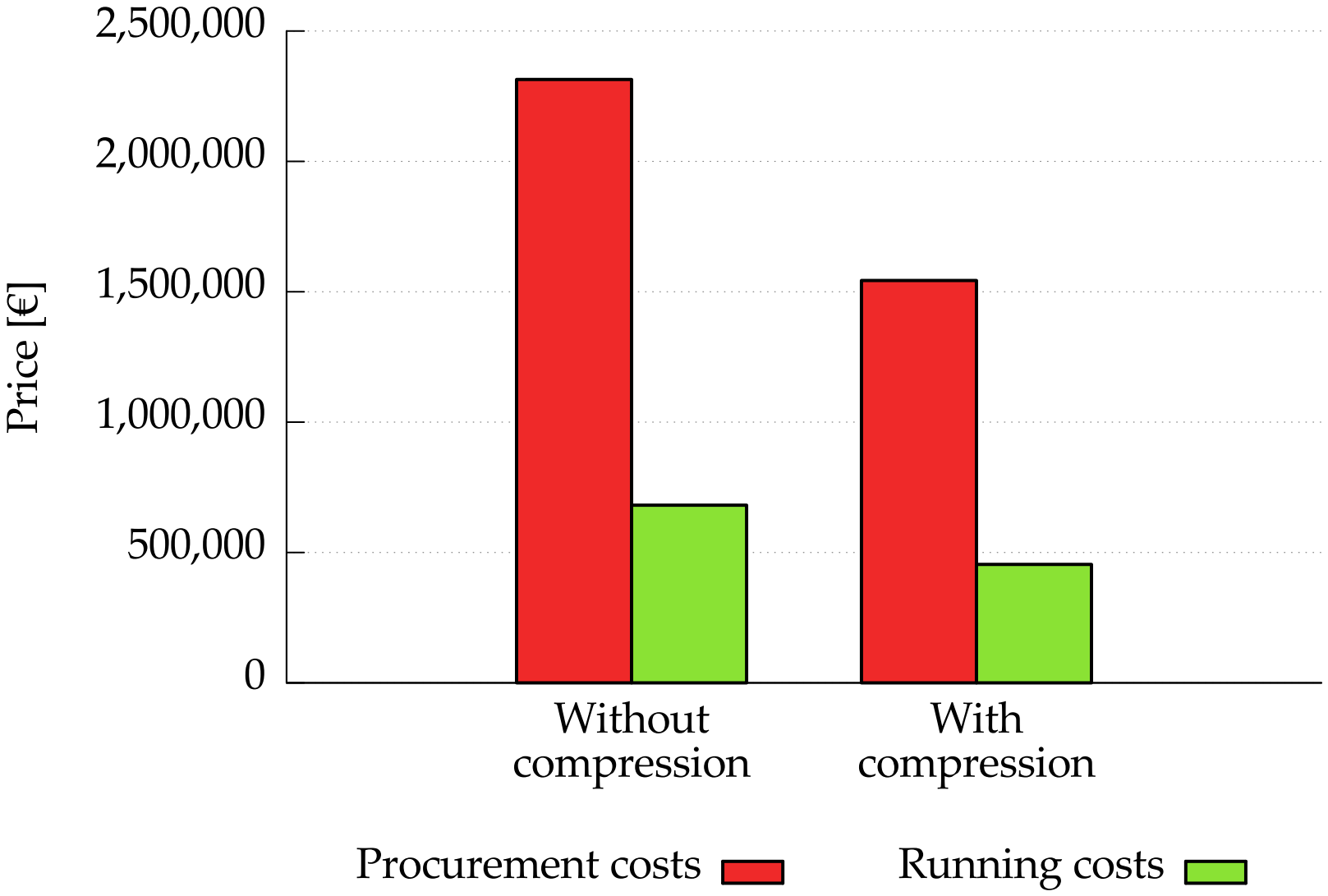
- TCO can be split up into two parts
  - Procurement costs:

$$\left\lceil \frac{n_{disks}}{ratio} \right\rceil \cdot c_{disk} + \left\lceil \frac{n_{disks}}{n_{dps} \cdot ratio} \right\rceil \cdot c_{server}$$

  - Running costs:

$$\left(\left\lceil \frac{n_{disks}}{ratio} \right\rceil \cdot p_{disk} + \left\lceil \frac{n_{disks}}{n_{dps} \cdot ratio} \right\rceil \cdot p_{server} \right) \cdot t_{system} \cdot c_{energy}$$

- Systems without compression use $ratio = 1$
- p_{disk} and p_{server} for uncompressed and compressed cases (`lz4`)


2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
Price [€]
Without compression
With compression
Procurement costs
Running costs

- Compression in HPC storage servers can be used to reduce costs
  - Compression should be turned on by default
  - High compression ratios directly correspond to high savings
- Carefully choose compression algorithms due to CPU overhead
  - `lz4` is a suitable compression algorithm for climate data
  - No significant increase in CPU utilization
- Improve the TCO model to take more factors into account
  - Different compression algorithms and indirect savings
- Benefits for different data sets have to be analyzed