Integrating I/O Measurement into Performance Optimisation and Productivity (POP) Metrics

PDSW 2019: 4TH INTERNATIONAL PARALLEL DATA SYSTEMS WORKSHOP

IT Center | RWTH AACHEN UNIVERSITY

Background

- **What is POP?**

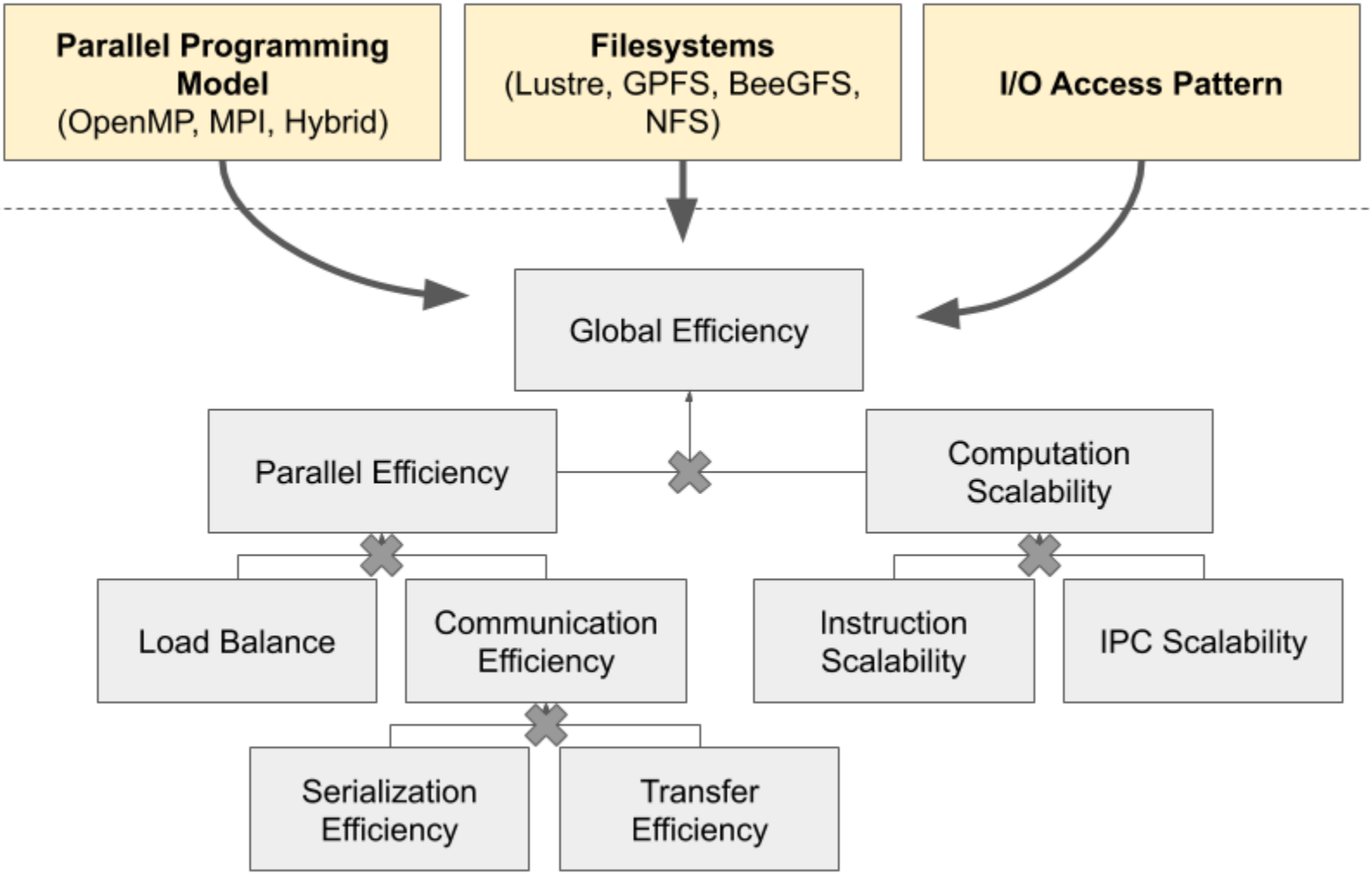
Center of Excellence that provides service to analyze parallel codes for academia and industry within the European Union to promote best practice in parallel programming.

The goal of the current POP metrics is to sort out components affecting performance in a way to make it easy to read and understand

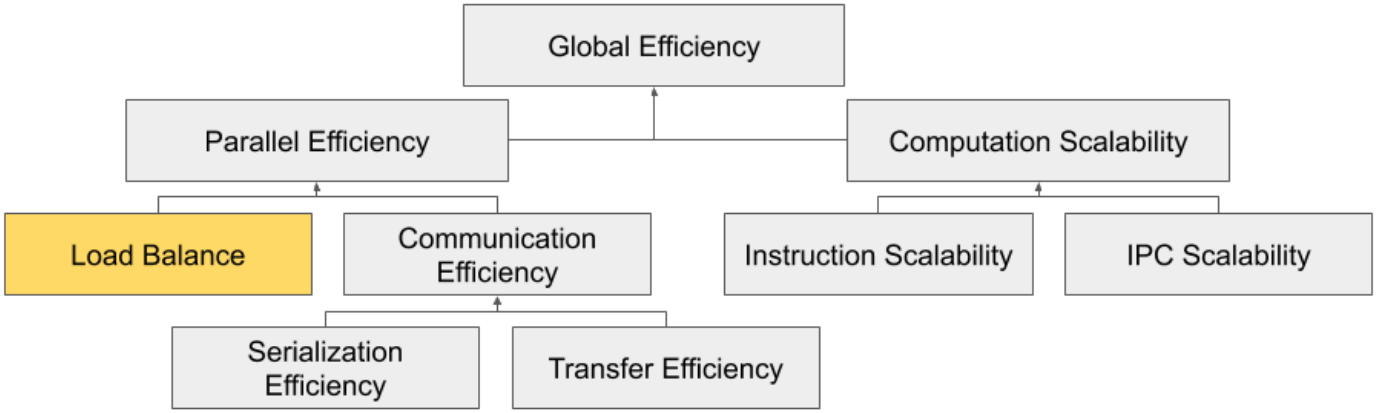
- **Unfortunately…**

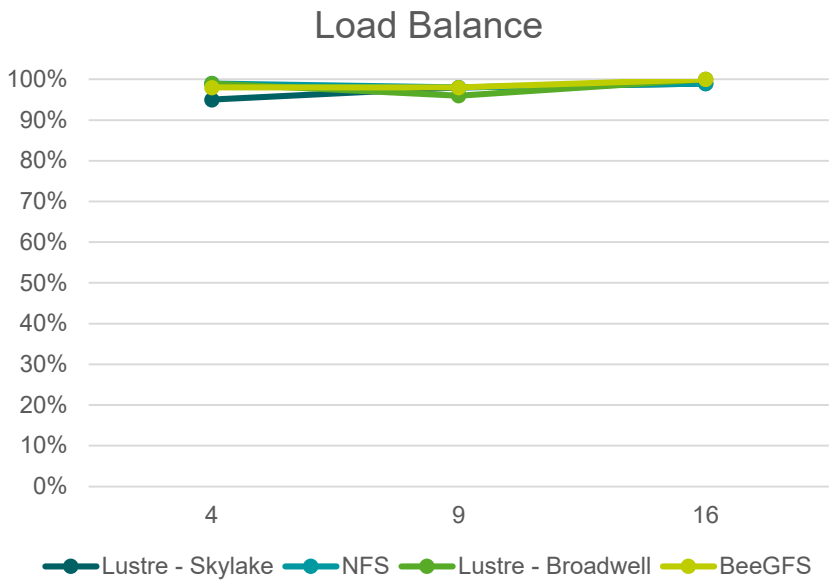
I/O was not considered inside this model yet

Methodology

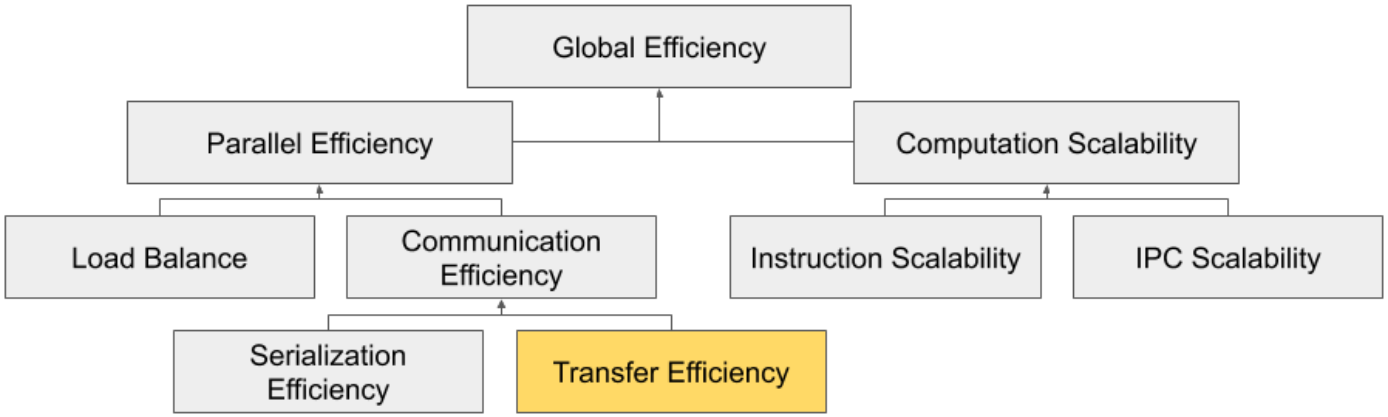


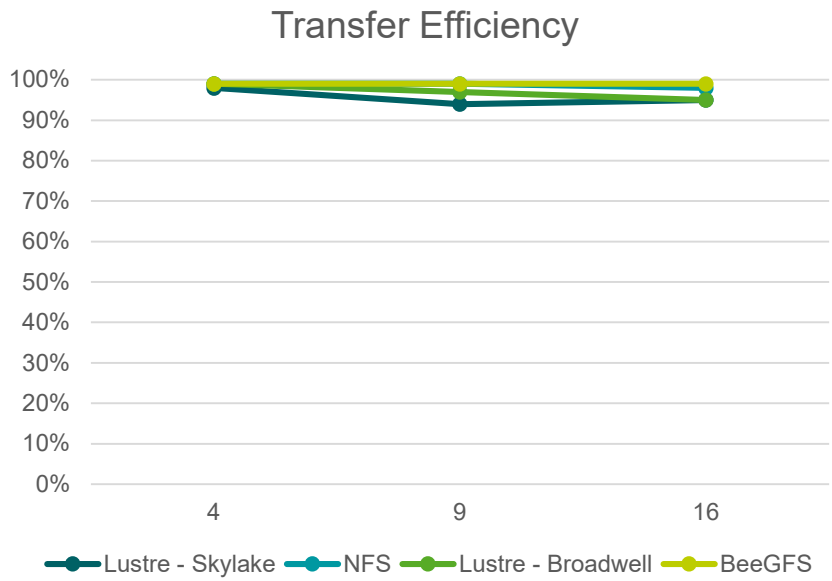
Current Impact on I/O Metrics with Collective IO Buffering (1)



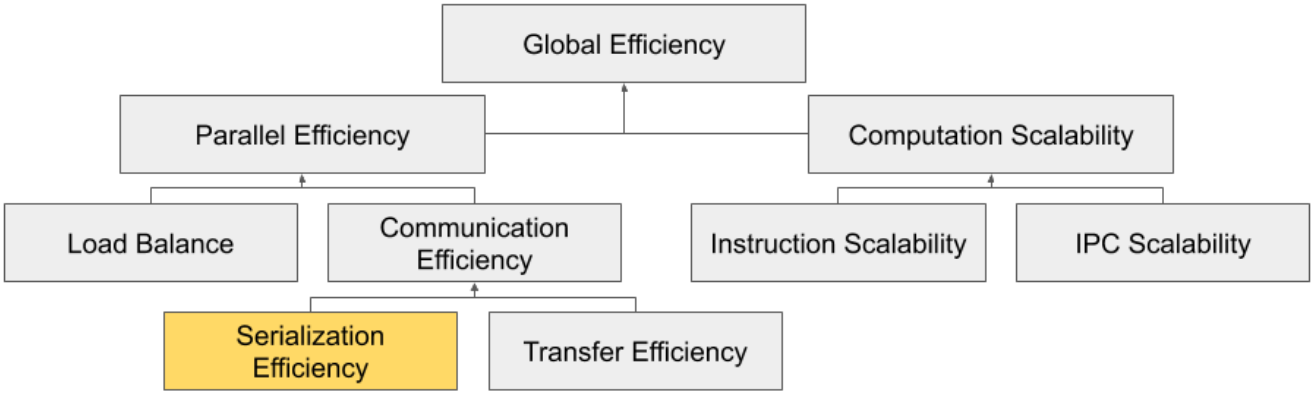


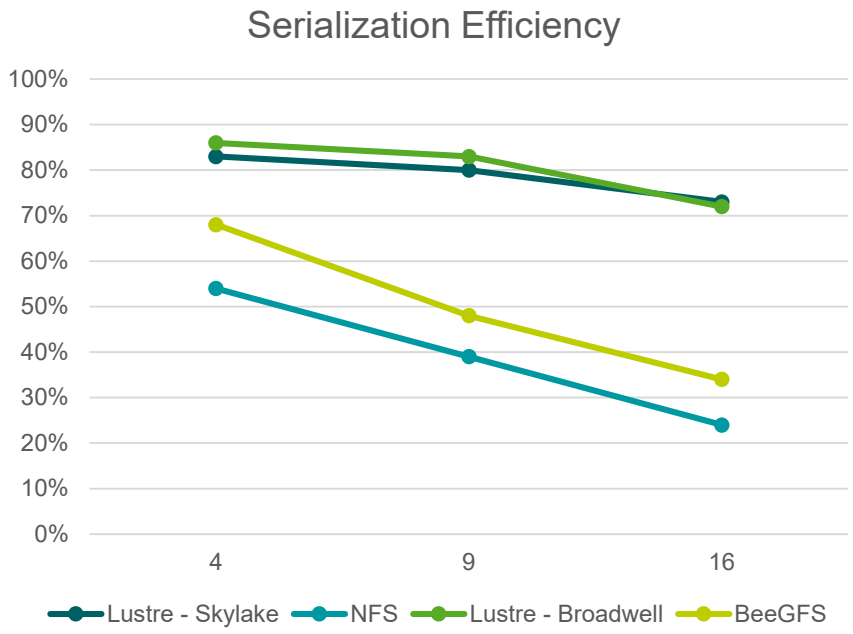
Current Impact on I/O Metrics with Collective IO Buffering (2)



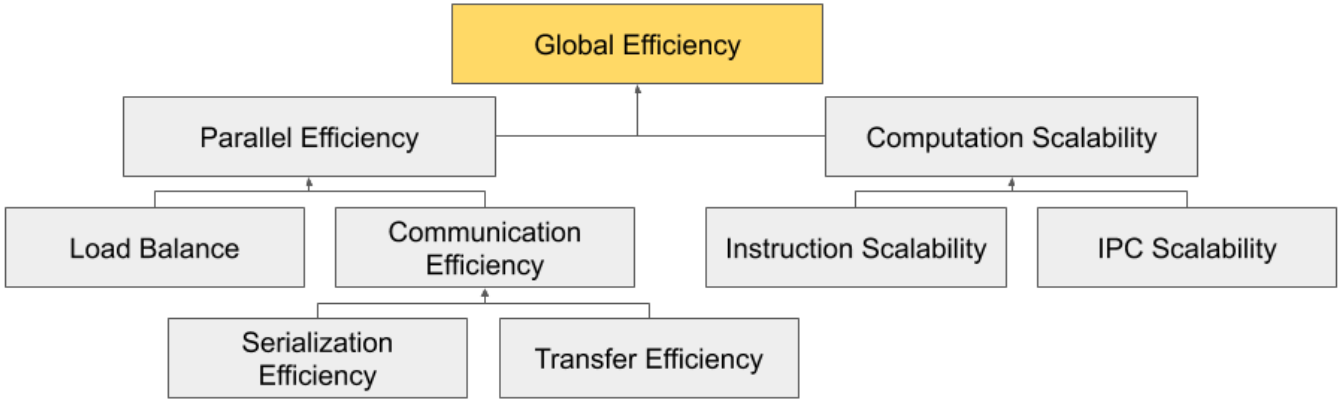


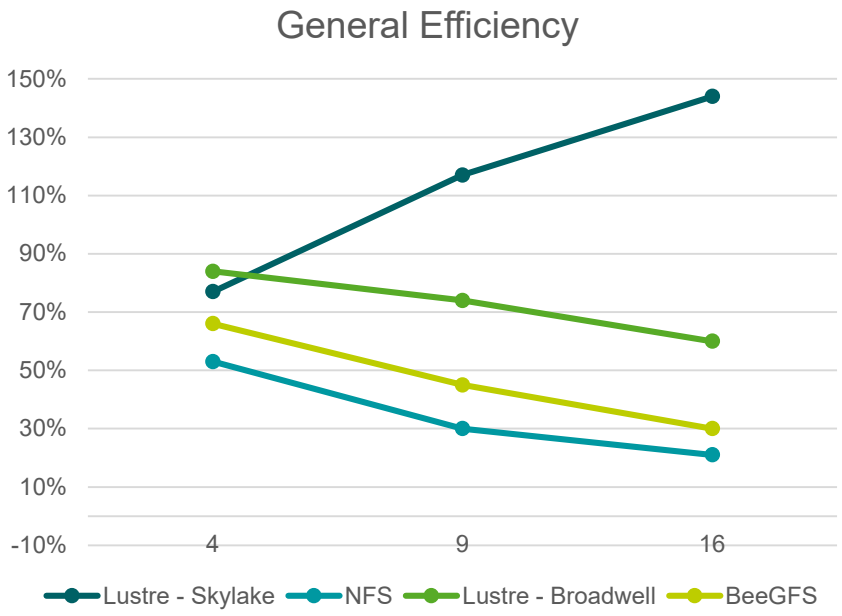
Current Impact on I/O Metrics with Collective IO Buffering (3)





Current Impact on I/O Metrics with Collective IO Buffering (4)





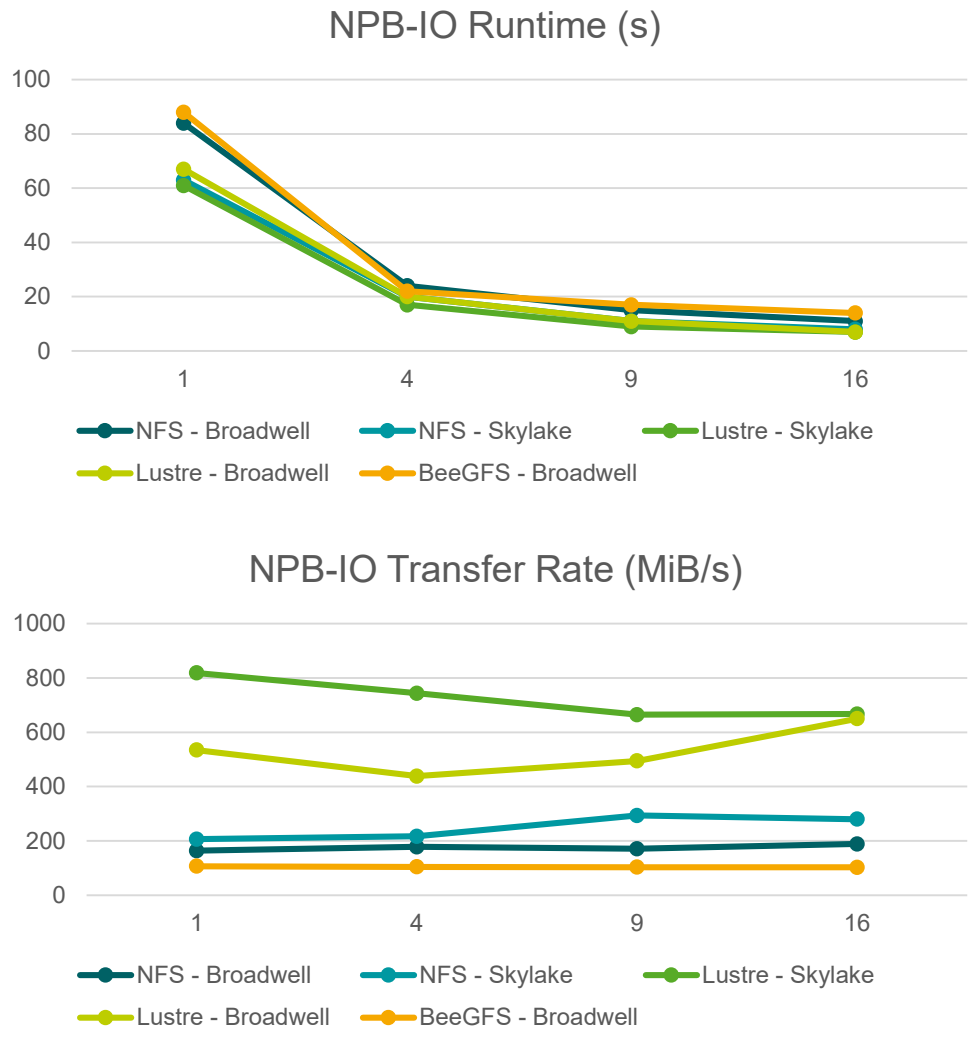
Initial Conculsion & Next Steps

- For MPI-IO with collective buffering case, the file systems difference appears on serialization efficiency this is due to:
  - I/O time is not evaluated on the ideal situation where I/O transfer rate is not a problem
- Performing more tests on various applications with different I/O size and pattern.
- Evaluating tools and methodologies to generate information that can represent the new I/O metric

Addendum

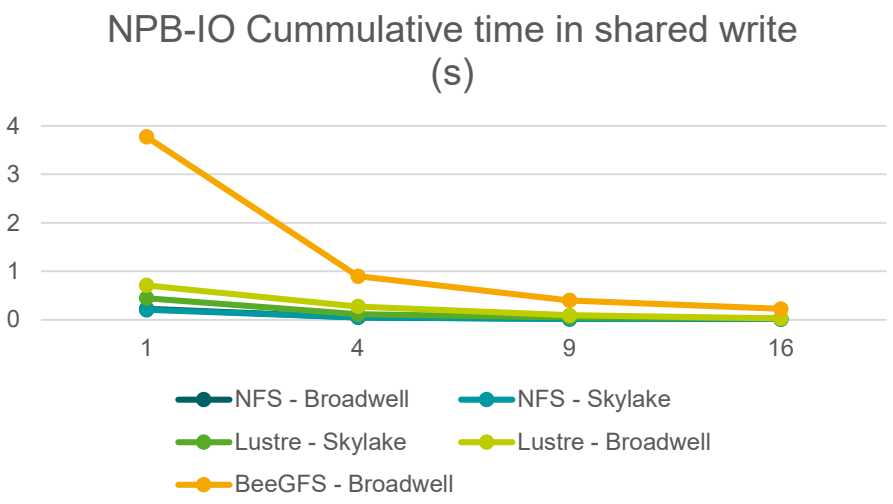
Additional result & Information

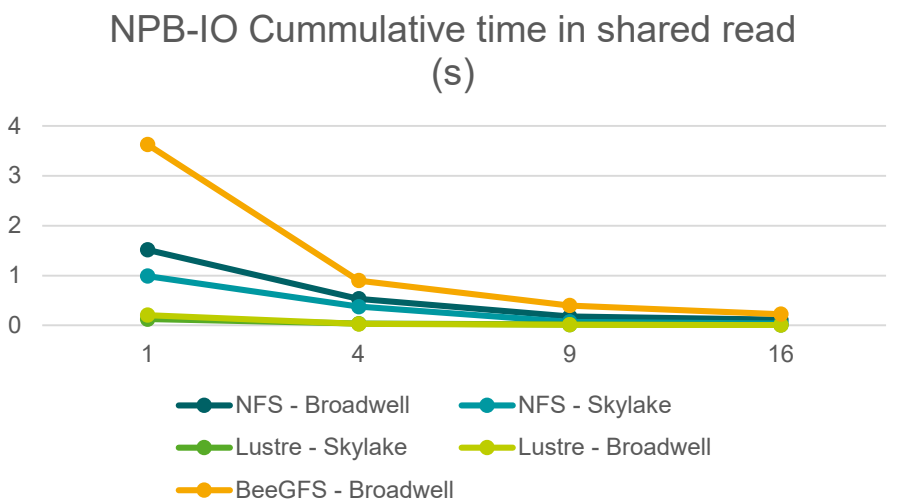
Darshan I/O result for NAS Parallel Benchmark (1)

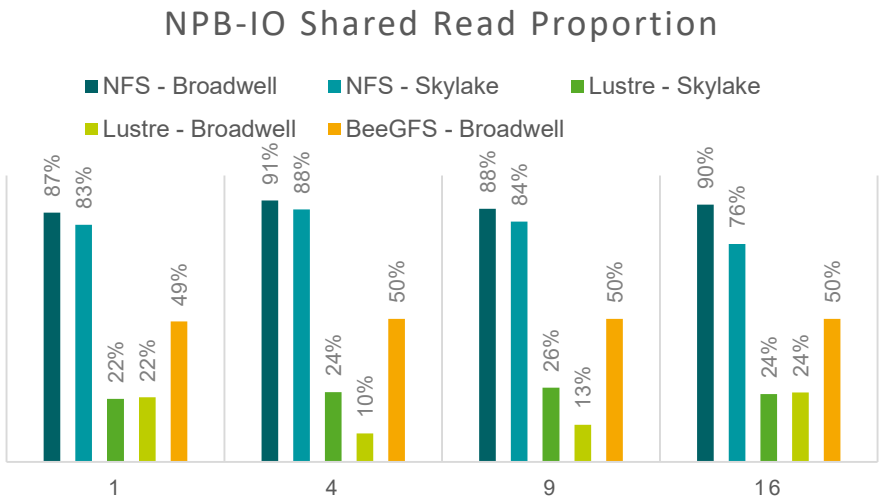


- Lustre filesystems both on Skylake & Broadwell has higher transfer rate than the other filesystem.
- This contributes to smaller runtime compared to the other filesystems.
- We can also see the impact on the compute cluster where Intel Skylake faster runtime

Darshan I/O result for NAS Parallel Benchmark (2)

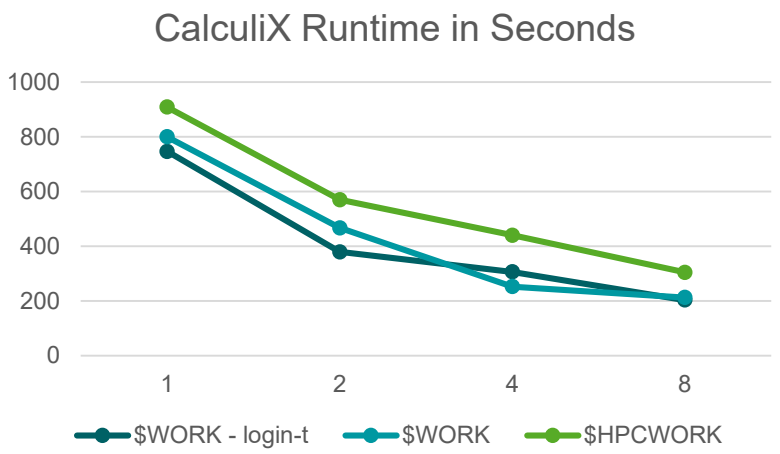




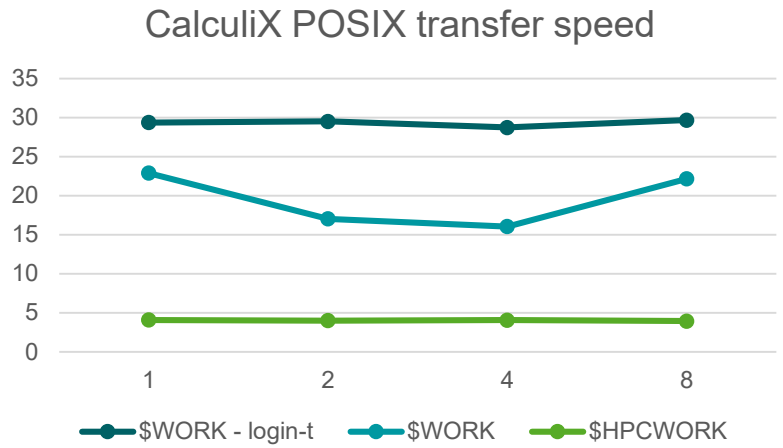


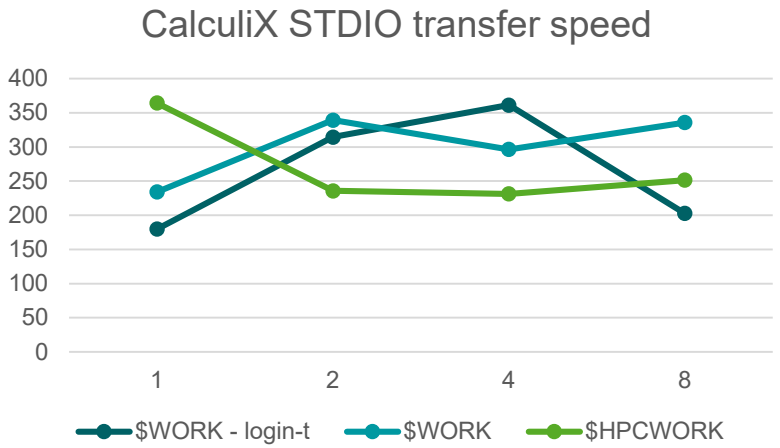
- Lustre shows good performance on reading file and not for writing
- BeeGFS shows balanced proportion for both read and write

CalculiX I/O result for NAS Parallel Benchmark (1)

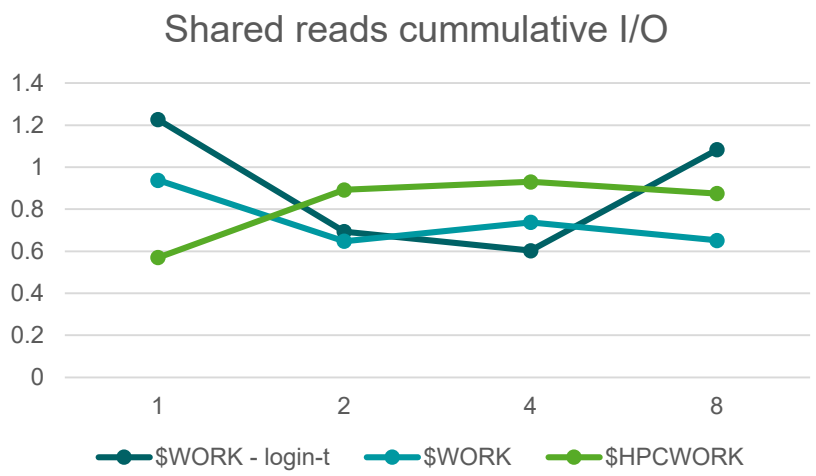


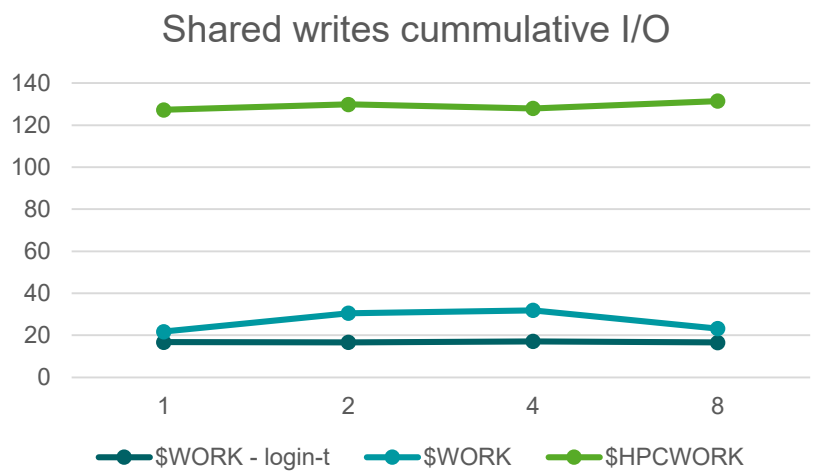
- Good efficiency based on POP metrics
- Lustre filesystem in the $HPCWORK performs worse than the other filesystem performance. Initial hypotheses: POSIX data transfer is mainly for writing and Lustry shared write performance is slower

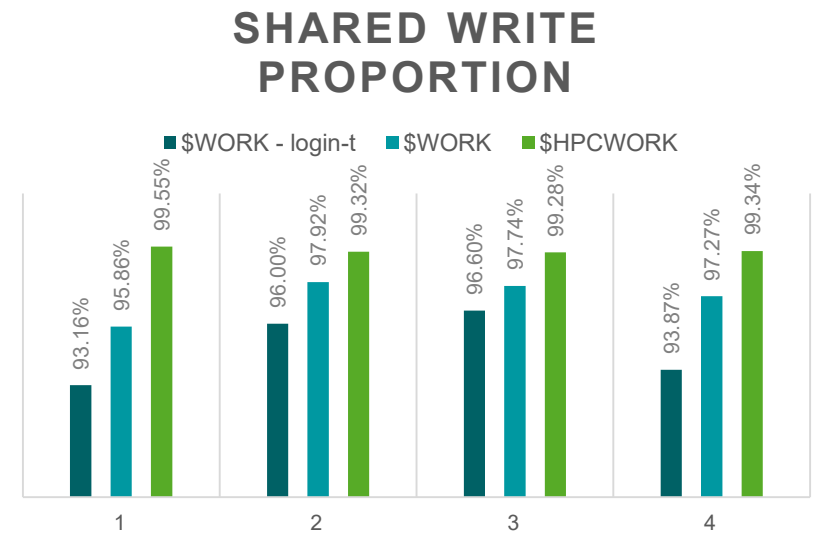




CalculiX I/O result for NAS Parallel Benchmark (2)







- Lustre performs badly doing file writing and CalculiX program creates and writes into 5 files continuously
- This is the case when the filesystem type affects the performance. In runtime result on the previous slide we can see that $HPCWORK result is the slowest among all three

Background

- Increased importance of the I/O optimization of the HPC application.
- The topic is challenging due to various moving variables that make measurement difficult.
  - Measuring I/O computation time within shared file systems needs to consider cluster workloads, filesystem type, and the chosen programming model
- POP is a Center of Excellence that provides service to analyze parallel codes for academia and industry within the European Union to promote best practice in parallel programming.
- The goal of the current POP metrics is to sort out components affecting performance in a way to make it easy to read and understand. The new I/O performance metrics should conform to this model

POP Metrics Explanation

- **General Efficiency Metric**
  Compound metric from parallel efficiency * computation efficiency
    - **Parallel Efficiency**
      compound metrics from load balance * communication efficiency
        - **Load Balance:** average computation time / maximum computation time
        - **Communication Efficiency:** maximum computation time / total runtime
            - **Serialization Efficiency:**
              maximum computation time on ideal network / total runtime on ideal network
            - **Transfer Efficiency:**
              total runtime on ideal network / total runtime on real network
    - **Computation Efficiency**
      ratios of total time in useful computation summed over all processes.

Source: https://pop-coe.eu/node/69

Test Case Environment

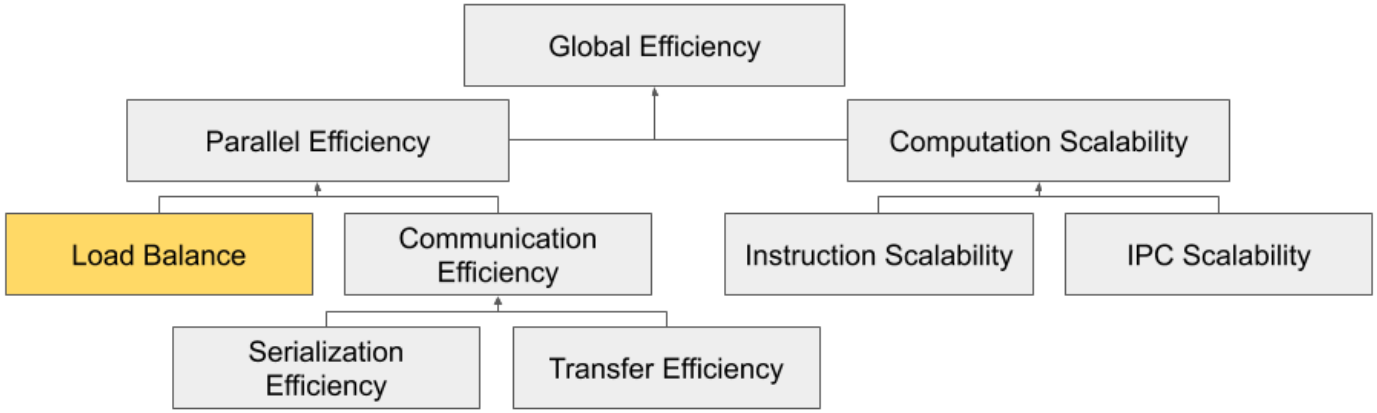
Software Information:

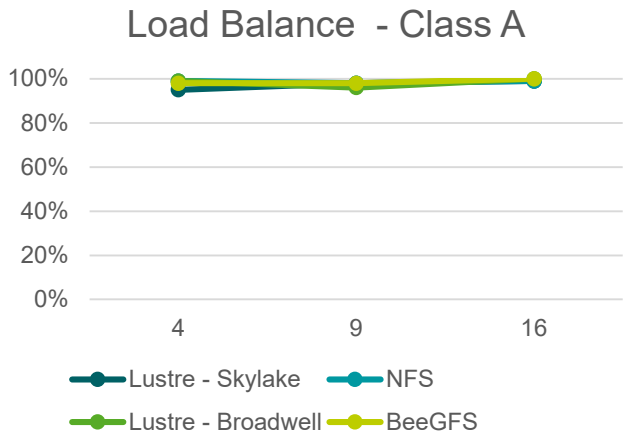
- NAS Parallel Benchmark
  - Subtype full: MPI I/O with collective buffering
  - Size A, B, C
  - Compiled with Intel compiler 2018.4
- CalculiX
  - Open source finite state element analysis application
  - POSIX I/O
  - Compiled with Intel compiler 2018.4

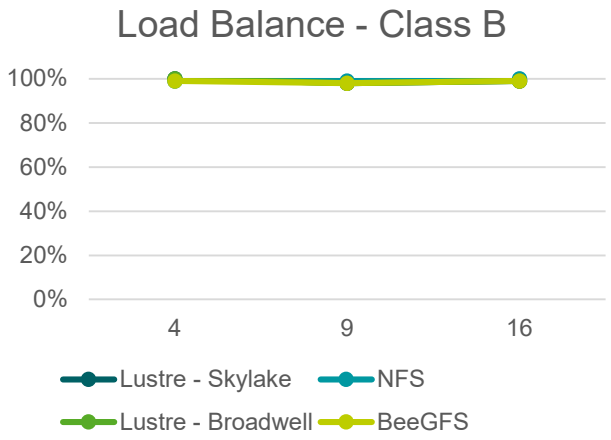
Hardware:

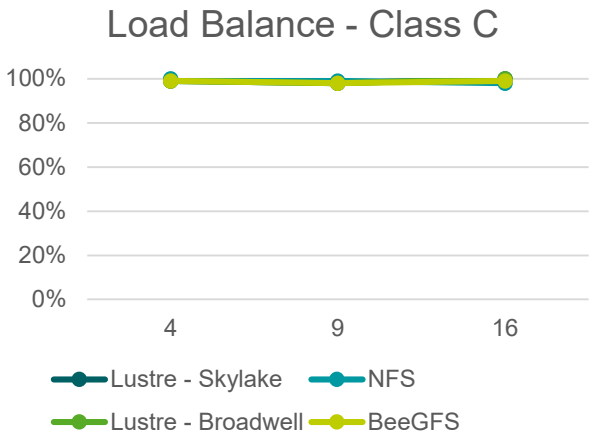
- RWTH Aachen University CLAIX18 compute cluster
  - Intel Skylake
  - Filesystems: NFS, Lustre
- RWTH Aachen University CLAIX16 compute cluster
  - Intel Broadwell
  - Filesystems: NFS, Lustre, BeeGFS

Current Impact on I/O Metrics (1)

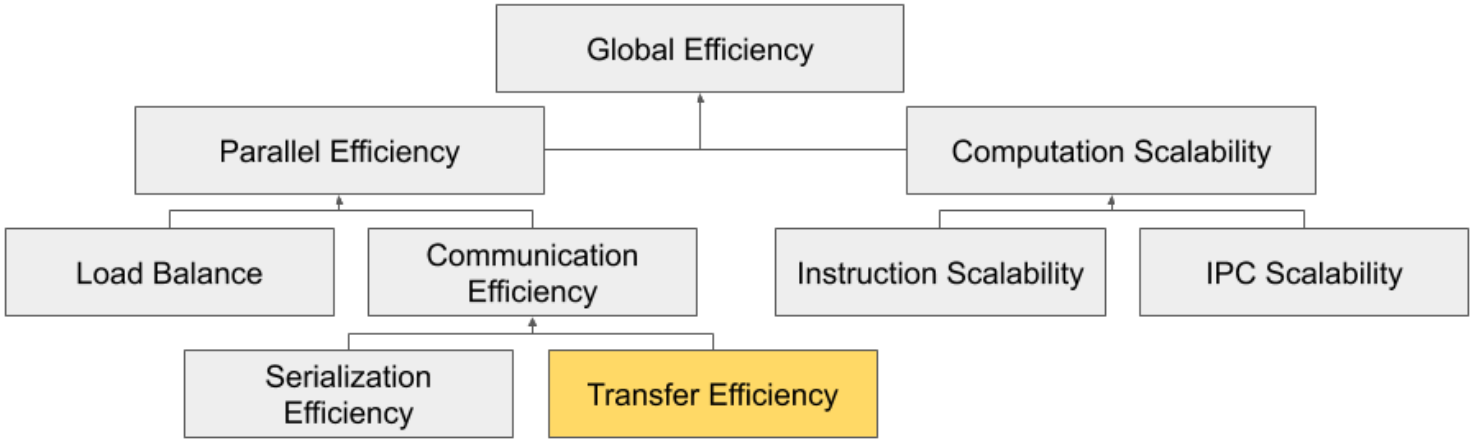


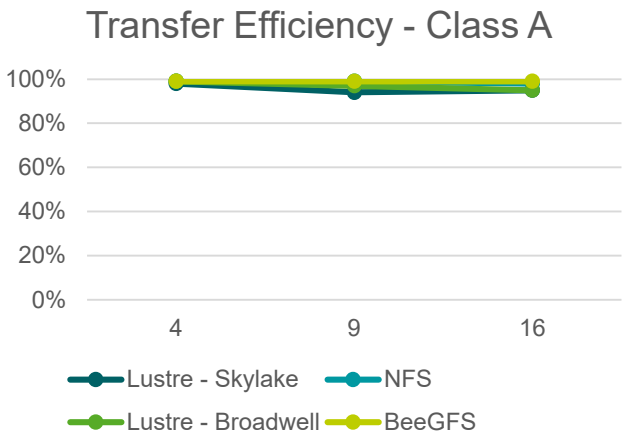


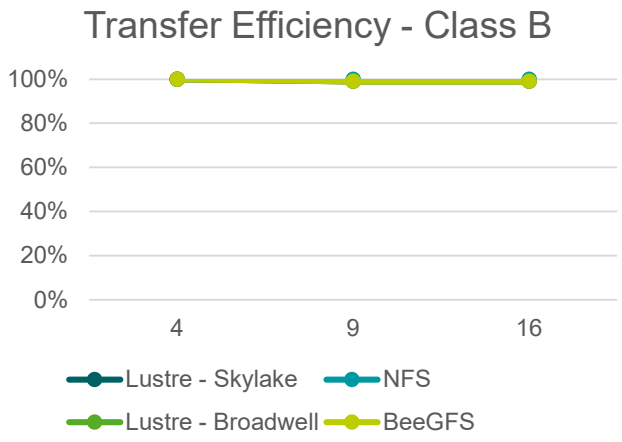


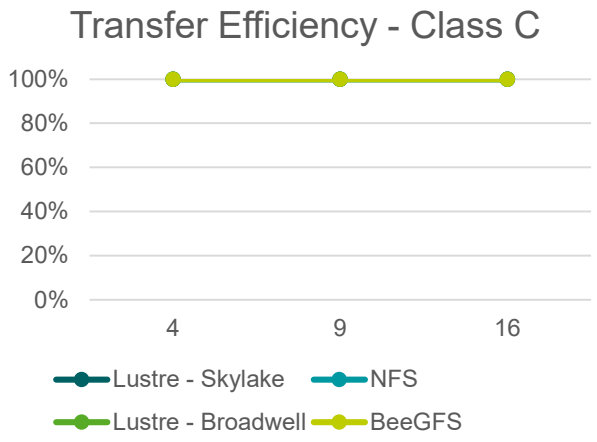


Current Impact on I/O Metrics (2)

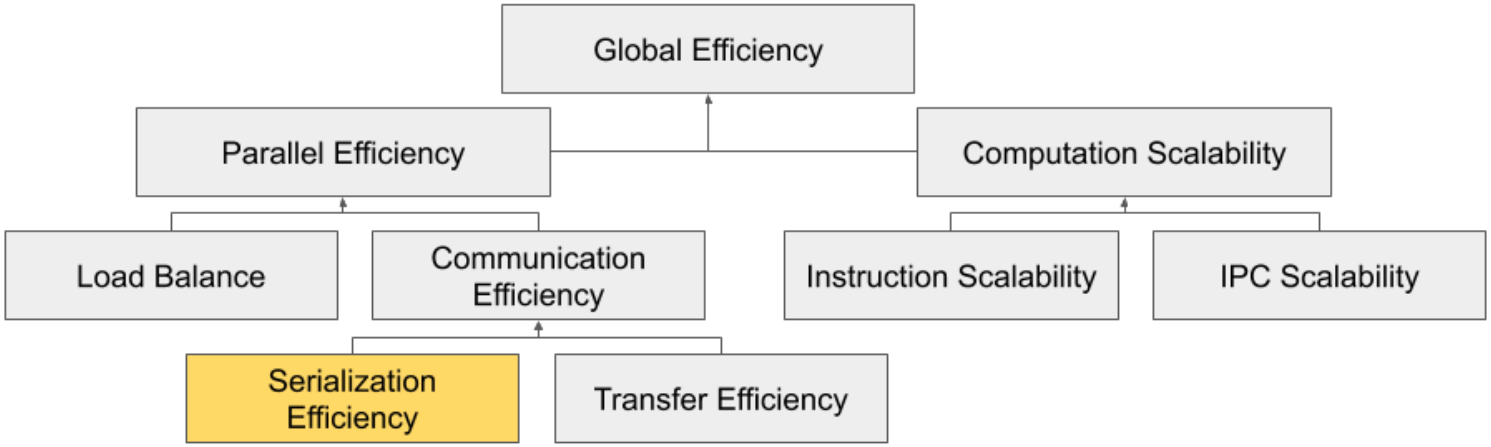


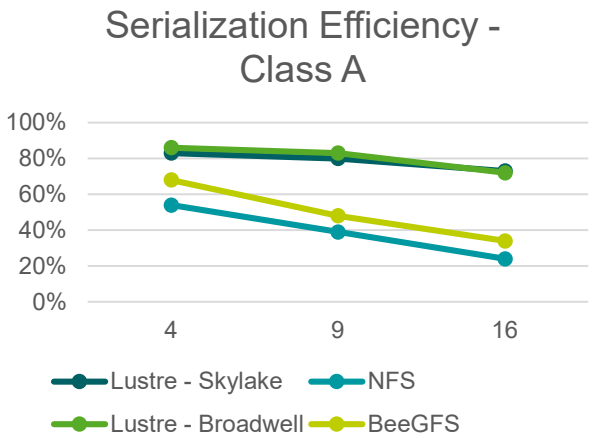


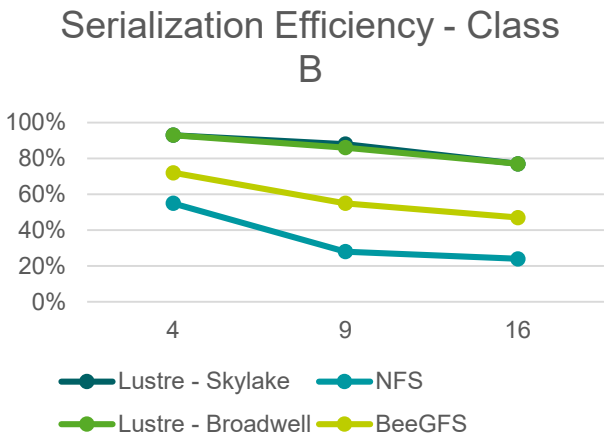


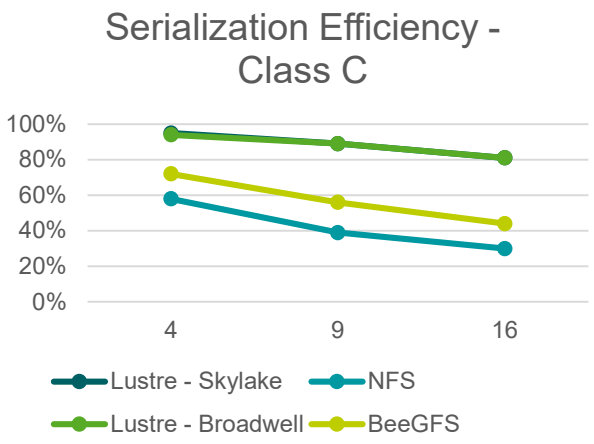


Current Impact on I/O Metrics (3)









Current Impact on I/O Metrics (4)