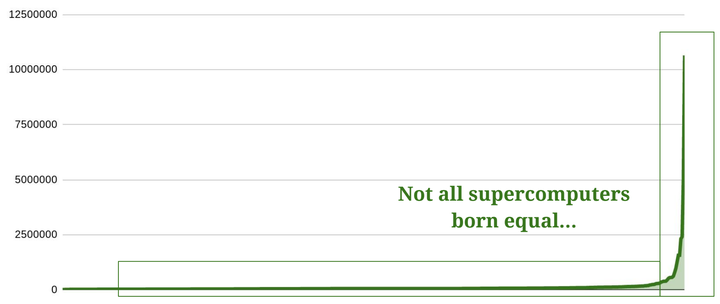
# Initial Evaluation of InfiniBox® Storage System For Embarrassingly Parallel HPC Applications In Small to Mid-range Supercomputers

Harel Levin[1], Gal Oren[1,2], Ilan Mizrahi[1], Emil Malka[1], Eran Brown[3]

[1] Department of Physics, Nuclear Research Center - Negev, P.O.B. 9001, Be'er-Sheva, Israel.
[2] Department of Computer Science, Ben-Gurion University of the Negev, P.O.B. 653, Be'er Sheva, Israel.
[3] Infinidat Inc., P.O.B. 46725, Herzliya Pituach, Israel.

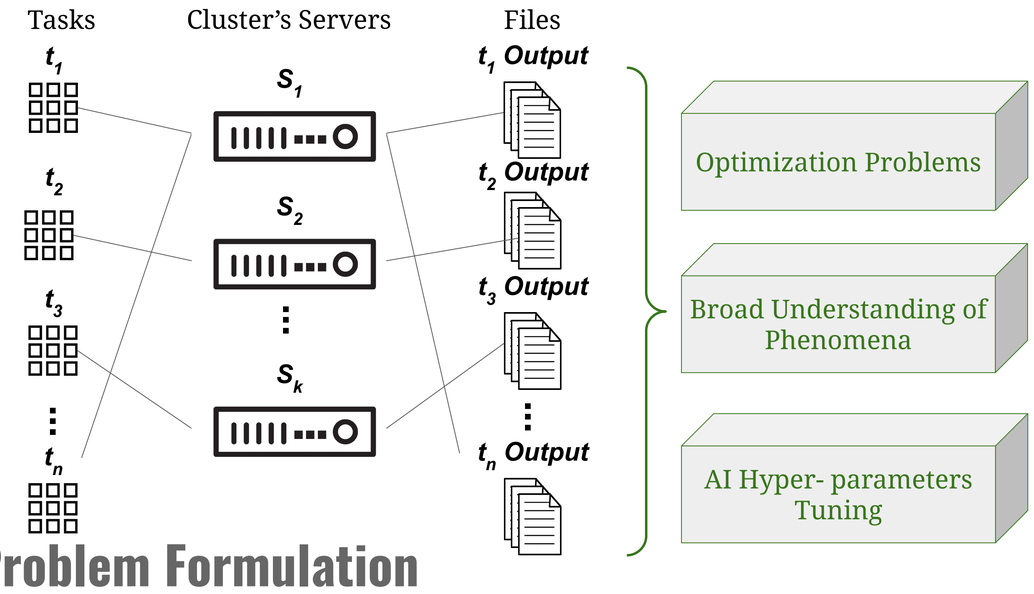
## Introduction

- Ever since the Bewoulf supercomputer back in 1994, **Commodity of-the-shelf High Performance Computers (COTS HPCs) had tremendously wide-spreaded**. These clusters make up a disproportionate number of systems deployed at scientific research institutions.
- As demonstrated in the chart below, while the top-notch of the supercomputers are consist of millions of cores, **most of the daily used clusters are made out of much modest amounts of compute nodes**.




**Top 500 (November 2019): Amount of Cores by Supercomputer**

- Over the last couple of years, HPC clusters are increasingly utilized for **embarrassingly parallel applications** for parameters surveys. The figure below demonstrates the fashion of these applications (on the left side) and several of common use-cases (on the right side).
- This kind of execution ends up in **massive amount of independent, relatively small files**. Furthermore, both read-from and write-to these files is often **sequential**.




## Problem Formulation

There are two common types of storage solutions integrated in HPC clusters. Distributed File Systems (such as Lustre and GPFS), and commodity NAS machines (such as NetApp and EMC).

### Distributed File Systems

**Pros**

+ **High throughput.**
+ Scalability.

**Cons**

- **Poor performance for random I/O to large amount of small files.**
- Expensive.
- Complicated deployment.
- Hard maintenance.

### Commodity NAS Storage Machines

**Pros**

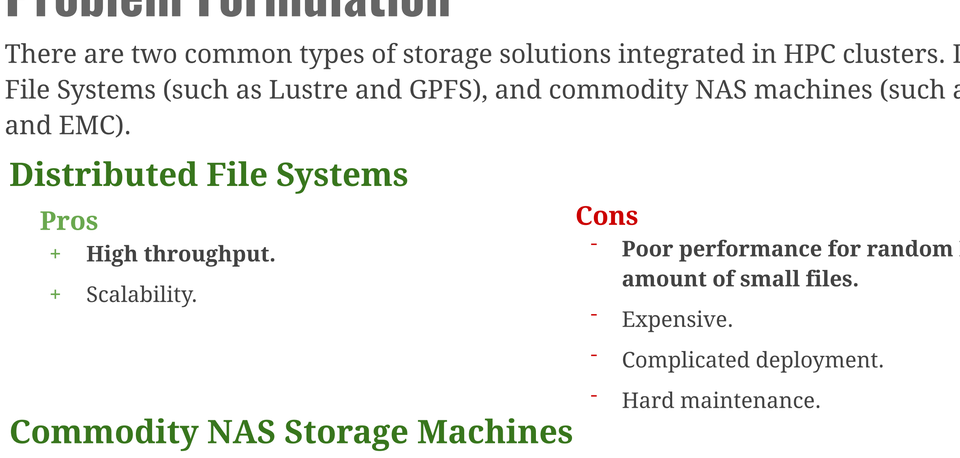
+ **Cheap.**
+ Easy Plug-and-Play deployment.
+ Easy maintenance.

**Cons**

- **Poor performance for parallel I/O.**
- Low Throughput.
- Single (or Double, at mose) Point of Failure.

The differences between these two kinds of **storage solutions forms a gap as both of them does not fit the parameter survey scenario**. The traditional distributed file systems perform poorly due to the large amount of files, while the commodity NAS storage machines does not scale.

## Current Storage Solutions




## IO500 Benchmark Suite

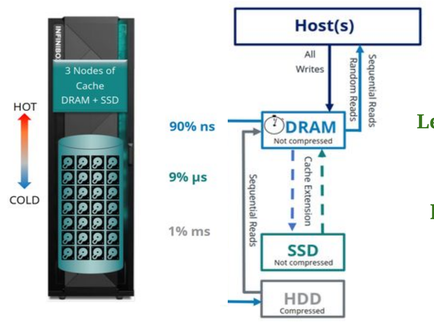
The IO500 benchmark suite was presented in ISC 2017. It consists of several I/O patterns.

- **IOR vs Mdtests** - IOR measures the bandwidth and mdtest measures the metadata.
- **Easy vs Hard** - Easy tests are optimized I/O patterns and Hard tests provides more challenging patterns.
- **Operation types** - Read, Write, Stat and Delete.

The amount of processes involved on the benchmark is predefined by the user.

## Infinidat® InfiniBox® Storage System

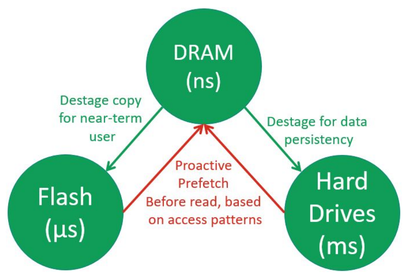
- InfiniBox® provides a **parallel multi-tiered storage system** with up to 12 fast Ethernet interfaces with up to 25Gb/s each (as for current Ethernet technology).
- Since **getting data from different tiers derives an overhead of 3 orders of magnitude in latency**, smart caching algorithms are required. At the same time, **minimizing the amount of data that needs to be placed in the expensive fast media enables reducing its size** (and therefore cost).




**Intelligent I/O Flow**

**Left: Infinidat® InfiniBox® architecture**

**Right: Data movement between DRAM, Flash & HDD mechanism**




- Towards this end, InfiniBox® includes a **group of algorithms collectively called 'Neural Cache'** that keep a connected history of every data section on the system and uses visibility into network traffic to the InfiniBox® to **find patterns in data access and predict future I/O requests**, which are then staged in cache for faster responses. By doing so, **InfiniBox® adjusts itself to any I/O pattern**.


**SSD vs. RAM Read Cache Hit Ratio using InfiniBox® at high load**

- **InfiniBox®'s 'Neural Cache'** handles data placement in real-time, enabling over 80-95% of reads from DRAM (See Green in chart above), and the majority of the rest to come from Flash (See Cyan in the chart above). This results in **low latency solution while keeping the majority of data on hard disks** which results in an optimal cost structure for large datasets.

## Performance Analysis

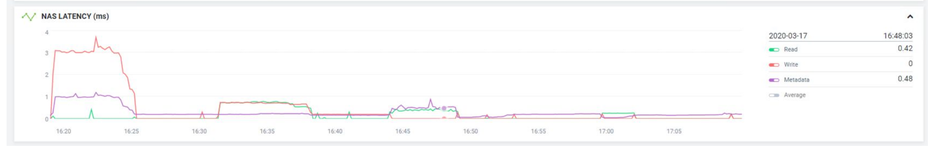
Our Evaluations were done using 5 out of InfiniBox®'s 12 network interfaces. Each interface served up to 5 processes simultaneously. InfiniBox®'s evaluations presents:

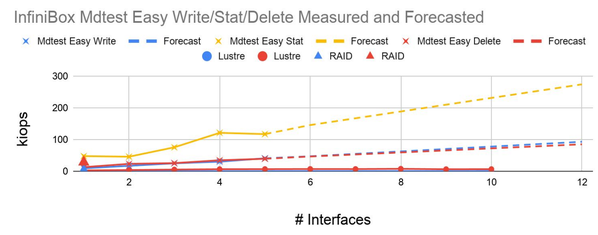
- **Linear scale-up for the independent (easy) tests, both for metadata operations and data throughput, which delivers good specifications for parameters surveys.**
- **Linear, yet suboptimal, results for hard patterns (similarly to other DFS systems).**

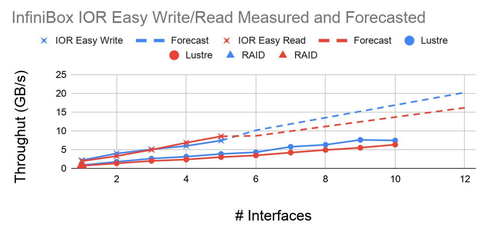
**Mdtest:**

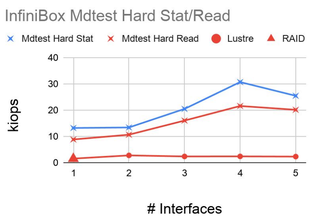

**IOR:**

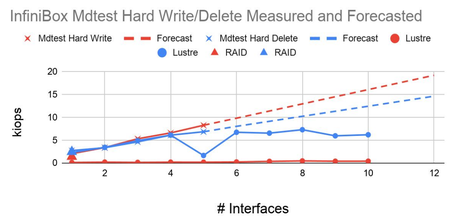

**Latency < 4ms**

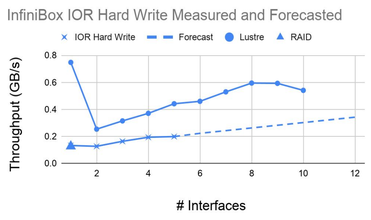



**Results of classic IO500 benchmark tests as was measured by InfiniBox® machine using 5 network interfaces (top) and by the IO500 application using 1 to 5 interfaces on InfiniBox®, 1 to 10 56Gb/s interfaces on Lustre filesystem, and a single interface on a commodity RAID storage machine (bottom). Notice that the RAID is inherently not scalable.**

## Conclusions & Future Directions

- **InfiniDat® InfiniBox® is a novel multi-tiered storage solution presents cutting-edge machine learning algorithms for a real-time cache-policy adjustment which decreases the overall latency of the system dramatically.**
- Our evaluations showed that InfiniBox® is a **perfect match for embarrassingly-parallel scenarios on mid-range HPC systems**, such as the highly common parameter surveys.
- A future work will look for software-level optimizations which will lead to superior performance for hard scenarios.

**Acknowledgments:** This work was supported by the Lynn and William Frankel Center for Computer Science. Computational support was provided by the *NegevHPC project*.