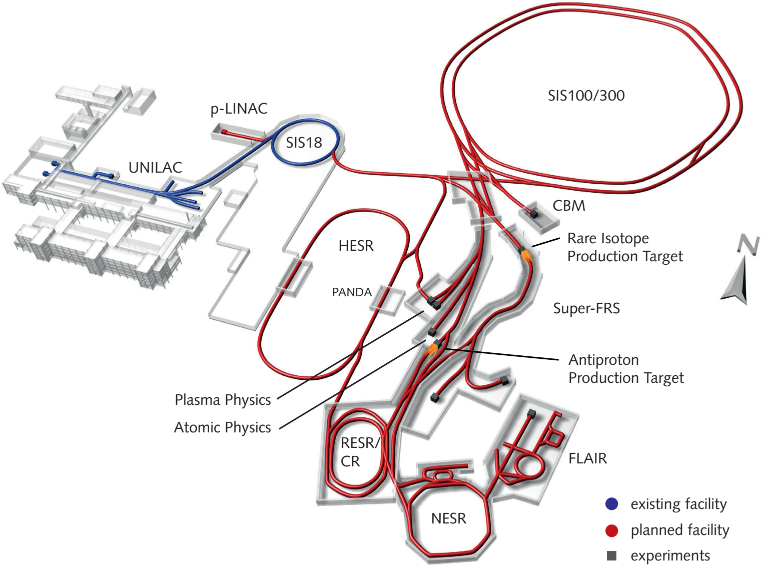
Storing and Analyzing Efficiently Big Data at GSI/FAIR

Thomas Stibor

GSI Helmholtz Centre for Heavy Ion Research, HPC

8. Mai 2014

Overview GSI/FAIR



I/O Requirements for FAIR

New concept:

- No hardware trigger

Flexible Event Selector:

- Compute farm calculates "trigger", ca. 60 000 cores only for CBM first level event selector

I/O:

- $\approx$ 1 TByte/sec for Compressed Baryonic Matter (short CBM)
- $\approx$ 1/2 TByte/sec for Anti-Proton Annihilation at Darmstadt (short PANDA)
- additional "smaller" Experiments

I/O Requirements for FAIR (cont.)

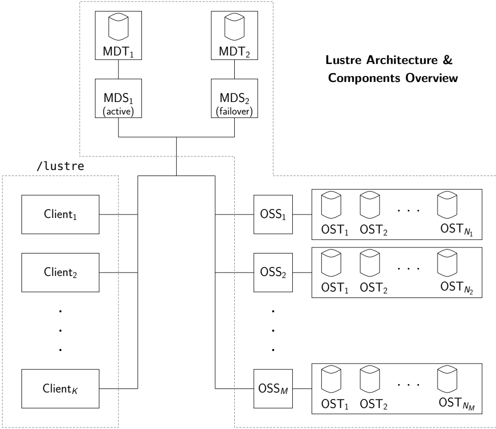
I/O after first level event selector:

- 1 GByte/sec for CBM
- 1 GByte/sec for PANDA
- additional "smaller" Experiments

In summary: Massive amount of data needs to be processed and stored.

Lustre Overview

- Lustre is a parallel distributed network file system for the domain of HPC
- 70% of the TOP500 supercomputers run Lustre
- POSIX compliant
- Lustre is free software (GPL v2)



Lustre Architecture & Components Overview

Lustre Deployment at GSI

Current:

- 7000 Disks.
- Raw capacity: 6.2 PByte
- Clients to OSS's: 1.5 TBit/sec I/O.
- OSS's to OSS's: 2.4 TBit/sec I/O.

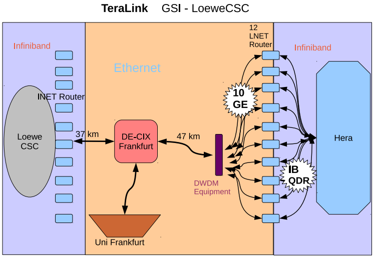
Exploring and Testing:

- Lustre 2.5 with ZFS back-end file system (software RAID).
- Multiple meta-data (MDT) servers (parallelize meta-data performance).
- Modeling and predicting file system "behavior" with probabilistic graphical models.

Highspeed Connections to Partnered Institutes

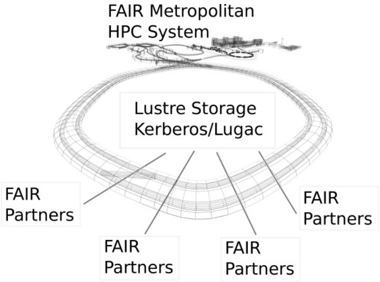
Current:

- 12×10 GBit/sec Ethernet with LNET/Lustre.
- Full bandwidth saturation.



Long term goal:

- LU-2221 ptlrpc: kerberos support for kernel >= 2.6.24
- LU-2392 kerberos: GSS keyring is broken >= 2.6.29
- LU-2384 kerberos: Support for MIT-kerberos >= 1.8.X is broken
- ...

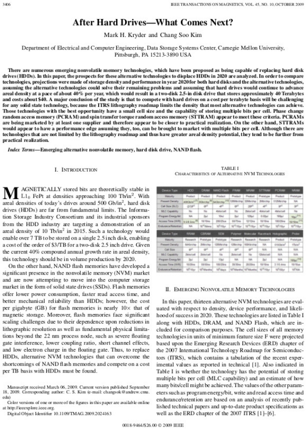


On Big Data

Kryder's Law:

> *The density of hard drives increases by a factor of* 1 000 *every* 10.5 *years (doubling every 13 months).*

source: http://en.wikipedia.org/wiki/Mark_Kryder



After Hard Drives—What Comes Next?

Mark H. Kryder and Chang Soo Kim

Department of Electrical and Computer Engineering, Data Storage Systems Center, Carnegie Mellon University, Pittsburgh, PA 15213-3890 USA

There are numerous emerging nonvolatile memory technologies, which have been proposed as being capable of replacing hard disk drives (HDDs). In this paper, the prospects for these alternative technologies to displace HDDs in 2020 are analyzed. In order to compare technologies, projections were made of storage density and performance in year 2020 for both hard disks and the alternative technologies, assuming the alternative technologies could solve their remaining problems and assuming that hard drives would continue to advance areal density at a pace of about 40% per year, which would result in a two-disk 2.5-in disk drive that stores approximately 40 Terabytes and costs about $40. A major conclusion of the study is that to compete with hard drives on a cost per terabyte basis will be challenging for any solid state technology, because the ITRS lithography roadmap limits the density that most alternative technologies can achieve. Those technologies with the best opportunity have a small cell size and the capability of storing multiple bits per cell. Phase change random access memory (PCRAM) and spin transfer torque random access memory (STTRAM) appear to meet these criteria. PCRAMs are being marketed by at least one supplier and therefore appear to be closer to practical realization. On the other hand, STTRAMs would appear to have a performance edge assuming they, too, can be brought to market with multiple bits per cell. Although there are technologies that are not limited by the lithography roadmap and thus have greater areal density potential, they tend to be further from practical realization.

Index Terms—**Emerging alternative nonvolatile memory, hard disk drive, NAND flash.**

I. Introduction

MAGNETICALLY stored bits are theoretically stable in $L1_0$ FePt at densities approaching 100 Tb/in^2. With areal densities of today's drives around 500 Gb/in^2, hard disk drives (HDDs) are far from fundamental limits. The Information Storage Industry Consortium and its industrial sponsors from the HDD industry are targeting a demonstration of an areal density of 10 Tb/in^2 in 2015. Such a technology would enable over 7 TB to be stored on a single 2.5 inch disk, enabling a cost of the order of $3/TB for a two-disk 2.5 inch drive. Given the current 40% compound annual growth rate in areal density, this technology should be in volume production by 2020.

On the other hand, NAND flash memories have developed a significant presence in the nonvolatile memory (NVM) market and are now attempting to move into the computer storage market in the form of solid state drives (SSDs). Flash memories offer lower power consumption, faster read access time, and better mechanical reliability than HDDs; however, the cost per gigabyte (GB) for flash memories is nearly 10× that of magnetic storage. Moreover, flash memories face significant scaling challenges due to their dependence upon reductions in lithographic resolution as well as fundamental physical limitations beyond the 22 nm process node, such as severe floating gate interference, lower coupling ratio, short channel effects, and low electron charge in the floating gate. Thus, to replace HDDs, alternative NVM technologies that can overcome the shortcomings of NAND flash memories and compete on a cost per TB basis with HDDs must be found.

Manuscript received March 06, 2009. Current version published September 18, 2009. Corresponding author: C. S. Kim (e-mail: changsok@andrew.cmu.edu)

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMAG.2009.2024163

TABLE I
Characteristics of Alternative NVM Technologies

Device Type	HDD	DRAM	NAND Flash	FRAM	MRAM	STTRAM	PCRAM	NRAM
Maturity	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Present Density	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Cell Size (SLC)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
MLC Capability	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Program Energy/bit	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Access Time (W/R)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Endurance/Retention	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Device Type	RRAM	CBRAM	SEM	Polymer	Molecular	Racetrack	Holographic	Probe
Maturity	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Present Density	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Cell Size	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
MLC Capability	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Program Energy/bit	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Access Time (W/R)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Endurance/Retention	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

II. Emerging Nonvolatile Memory Technologies

In this paper, thirteen alternative NVM technologies are evaluated with respect to density, device performance, and likelihood of success in 2020. These technologies are listed in Table I along with HDDs, DRAM, and NAND Flash, which are included for comparison purposes. The cell sizes of all memory technologies in units of minimum feature size F were projected based upon the Emerging Research Devices (ERD) chapter of the 2007 International Technology Roadmap for Semiconductors (ITRS), which contains a tabulation of the recent experimental values as reported in technical [1]. Also indicated in Table I is whether the technology has the potential of storing multiple bits per cell (MLC capability) and an estimate of how many bits/cell might be achieved. The values of the other parameters such as program energy/bit, write and read access time and endurance/retention are based on an analysis of recently published technical papers and up-to-date product specifications as well as the ERD chapter of the 2007 ITRS [1]–[6].

If hard drives continue to progress at their current pace, then in 2020 a two-disk 2.5-in disk drive can store approximately 40 Tera-Bytes and would cost about $40.

How about algorithms, are they also scale in such a manner, e.g. multiplying *big* matrices?

On Big Data and Matrix Multiplication

```
void mult_naive(double A[dim][dim], double B[dim][dim], double C[dim][dim], unsigned int dim)
{
    for (int i = 0; i < dim; i++) {
        for (int j = 0; j < dim; j++) {
            double sum = 0;
            for (int k = 0; k < dim; k++) {
                sum += A[i][k] * B[k][j];
            }
            C[i][j] = sum;
        }
    }
}
```

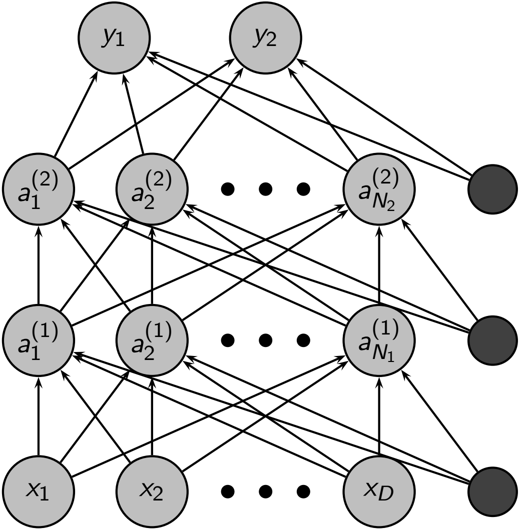
Run-time complexities big $\mathcal{O}$ notation for matrix multiplication algorithms:

$$
\begin{aligned}
\text{Naive example} &: \mathcal{O}(dim^3) \\
\text{Strassen(1969)} &: \mathcal{O}(dim^{2.8074}) \\
\text{Coppersmith–Winograd(1990)} &: \mathcal{O}(dim^{2.375477}) \\
\text{Francois Le Gall(2014)} &: \mathcal{O}(dim^{2.3728639})
\end{aligned}
$$

Consider large matrices, e.g. $10^5 \times 10^5$. How do we efficiently multiply those?

Matrix Multiplication in Neural Networks

Consider the problem of learning (deep) neural networks, which can be *perfectly* formulated in terms of matrix multiplications.



parameters to learn

$\mathbf{W}^{(3)}, \mathbf{b}^{(3)}$

$\mathbf{W}^{(2)}, \mathbf{b}^{(2)}$

$\mathbf{W}^{(1)}, \mathbf{b}^{(1)}$

Matrix Multiplication in Neural Networks (cont.)

- Activation of neuron: $a_1^{(1)} = f(\mathbf{W}_{1,1}^{(1)} x_1 + \mathbf{W}_{1,2}^{(1)} x_2 + \ldots + \mathbf{W}_{1,D}^{(1)} x_D + \mathbf{b}_1^{(1)})$, where $f(\cdot)$ is some activation function, e.g. $f(z) = 1/(1 + exp(-z))$.
- Forward-Pass (matrix multiplication, vector addition, element-wise activation function):

$$\mathbf{y} = f(\ldots f(\mathbf{W}^{(3)} f(\mathbf{W}^{(2)} f(\mathbf{W}^{(1)} \mathbf{X} + \mathbf{b}^{(1)}) + \mathbf{b}^{(2)}) + \mathbf{b}^{(3)}) \ldots)$$

- Backward-Pass for updating the parameters (weights) can also be formulated in terms of matrix multiplications.
- Forward-Pass $+$ Backward-Pass $\equiv$ Back-Propagation Algorithm.

Matrix Multiplication in Neural Networks (cont.)

- Training big and deep neural networks, e.g. 60 million parameters and 650 000 neurons and massive amount of data was infeasible 10 years ago.
- It took months to train big and deep neural networks, moreover one concluded that such neural networks are very prone to *overfitting* and the *gradient vanishing* problem.
- Since the revolution of powerful and cheap GPU's (and proper SDK), big and deep neural networks can be trained in a couple of hours or days.
- State of the art in computer vision (convolution neural network), speech recognition, natural language processing, etc..
- See work of: Geoffrey Hinton, Yann LeCun, Yoshua Bengio, Andrew Ng, and many more.

Summary

- When Kryder's Law still holds in the future, then we will be surrounded by massive amount of data.
- Highly optimized task specific GPU algorithms on very powerful GPU *can* enable to crunch this data.
- Algorithms processing and analyzing the data, ideally scale as well (online and parallized algorithms).