



The Need for High Throughput Storage



Terascale, Inc.
145 Bodwell St
Avon, MA
www.terascala.com

Today's users are faced with the dual challenges of delivering faster and more detailed results from their processing while at the same time working from larger and larger quantities of data.

Terascala builds high throughput, scalable storage appliances leveraging parallel file systems for customers who analyze large quantities of data to make faster, more accurate decisions.

Our solution delivers up excellent price / performance compared to alternative solutions with no application changes, offering existing users a better alternative while allowing new users to affordably run their applications

Understanding the Problem

Two key trends are, together, driving a growing need for high throughput, cost effective storage solutions. These trends are the rapid increase in the volume and types of data available and the growth in the use of commodity based servers to process that data.

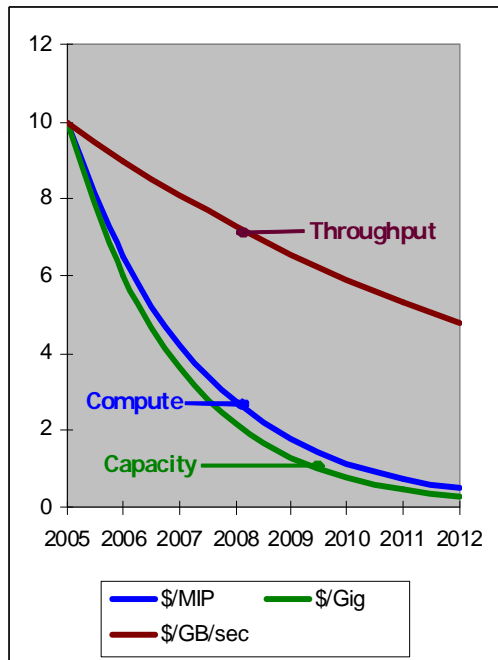
IDC predicts that the amount of data being stored will grow by over 10x (from 180 exabytes to 1800 exabytes) over the next 5 years. This data will include increasing volumes of the kinds of data historically stored, but, perhaps more importantly, will include significant growth in new types of data, including things like unstructured text data, videos, audio, and sensor based data. Many of these new data sources deliver data that is difficult to characterize and organize. Processing these new types of data requires new techniques and technologies.

Even as the data volume grows, users need to make more accurate, faster decisions in a cost effective manner. In many industries, this need / ability is changing old business models and creating new ones. In the oil and gas industry, users are leveraging more detailed seismic analysis to more accurately estimate and price future gas fields. For web based companies, the ability to quickly track, store and mine user behavior allows for better targeted advertising and content, increasing revenue. In financial services, the ability to quickly analyze multiple sources of structured market data combined with unstructured "news" accounts provides the ability to properly price investments in near real time. For companies and institutions across many industries, the ability to perform more detailed and faster simulations allows them to make better decisions. All these environments have the same key challenges in common, balancing the speed of computation, the speed of storage, and the total cost of solution to optimize their answers.

Every processing environment is made up of 3 key elements, compute, networking and storage. Over the past years, compute / processing technology has made great strides, as exemplified by Moore's law of the doubling of processing power every 2 years at the same price point. These strides in raw processing power have enabled many new application areas. As an example, the ability to perform computer simulations quickly and cost effectively have enabled many industries to develop new, better products rapidly without the time and expense of developing physical models.

As processing power has grown, the balance between networking and storage elements became lopsided. Many applications were limited by their ability to get data from outside sources to the processing elements themselves. The existing networking technologies could not keep up with the demands of the processors. This has driven newer, faster, more

processing focused networking technologies such as InfiniBand and 10 Gigabit Ethernet. These technologies have removed the network bottleneck, delivering the bandwidth and latency needed to keep the latest multi-core processing elements busy.



Normalized \$/metric

However, the improvements in processing and networking technologies have not been matched in the storage area. While storage capacity has improved at an incredible rate (exceeding even the processing “law” of 2x improvement every 2 years), storage throughput has not kept pace. Storage capacity has enabled users to capture and store tremendous amounts of data in a cost effective manner. Yet even as the amount of data stored increases, the ability to process that data into information that can be used quickly and efficiently has been limited.

Storage throughput, quantified as GB/sec, is the challenge of getting data from the disk to the application and is a function of the physical storage system, the file system running on it, the interconnect between the storage system and the compute nodes, and how well all those elements are optimized. The requirements for an ideal high throughput storage solution are the speed to keep the processing elements busy, the price / performance to enable broad deployment, and simplicity in installation and management.

Many users are faced with the choice of either a cost effective storage solution that is focused on capacity or an expensive, high throughput solution. The capacity focused storage systems are designed to cost effectively store massive amounts of data and, in many cases, allow for simple capacity scaling so that as the quantity of data stored increases, additional capacity can be added quickly and easily. The high throughput solutions tend to combine various complicated storage technologies (SANs, fibre channel, etc) to deliver the throughput needed but at a significant cost increase relative to the capacity provided while adding additional features not required for high throughput processing.

Many users find that they are limiting their overall capabilities to match the storage systems they have or can afford to purchase. Some users limit the size of their data sets to ensure that the storage system’s low throughput doesn’t limit the compute capabilities. Other users attempt to balance their environments by limiting the number of processing nodes that are leveraging a particular storage solution. Either of these solutions is potentially limiting the accuracy or detail of the results that they are getting. The last option is to invest in a high throughput storage solution, but, due to cost considerations, this is often used as a shared resource, limiting the number of users that can leverage this technology.

Terascala is changing this dynamic.

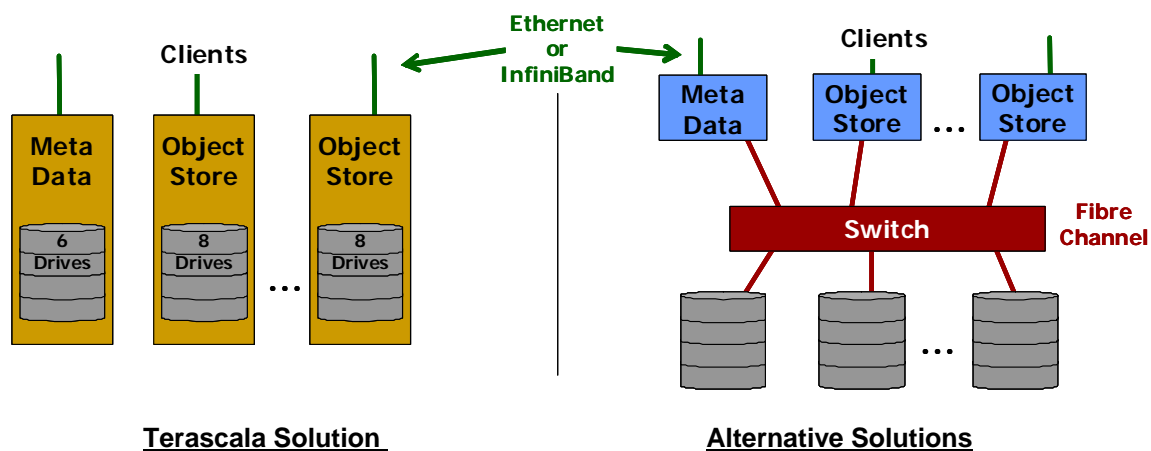
Terascale's Storage Appliance

The Terascale RTS 1000 Storage Appliance is a Lustre™ parallel file system based storage appliance offering high throughput and high capacity. It delivers excellent price / performance compared to alternative storage solutions.

The RTS 1000 is designed to deliver the maximum throughput to enable applications to run at peak efficiency. It can deliver over 2GB/sec from a single enclosure and up to 10 GB/sec for a full rack solution. Designed to plug directly into the compute client network environment, the RTS 1000 has an optimized data path from the client network through to the disks within the storage device.

Simplification of deployment and ongoing management is a key aspect of the RTS 1000. It is delivered as an appliance with all the software installed and tuned to deliver performance. With its built-in management system, it is easy to add additional capacity and throughput or to fine tune key aspects of the environment. The RTS 1000 is designed to be operated and managed by system administrators without extensive storage experience.

Terascale leverages Lustre because it is open source, high performance clustered storage software initially developed for applications needing very high throughput, scalability, and capacity. It offers high levels of reliability, scalability, and performance, having been deployed in some of the largest compute installations in the world. Lustre leverages a simple metadata / stored object architecture, where the metadata server stores location information about data and the object store servers act as the repository for the actual data. This approach allows throughput to be scaled by simply adding additional object store servers. At Terascale, we have optimized our architecture and tuned Lustre for the specific capabilities of the RTS 1000, so the appliance delivers a simple, easy to use and expand solution that can grow as your needs grow.



With its simplified architecture and use of commodity technology where possible, the RTS 1000 is able to deliver tremendous price / performance while simplifying the overall deployment and management of the solution. As shown in the figure below, the RTS 1000 solution has no fibre channel connections to add complexity and cost to its implementation.

Summary

As the volume and types of data being collected has continued to increase, users are being asked to leverage that information into faster, more accurate decisions. While compute power and storage capacity have improved both performance and cost effectiveness, storage throughput has not kept pace. The Terascale solution delivers the high throughput and scalability that users need in a cost effective appliance enabling them to leverage the most of their computing environments.

Contact Terascale to learn more about how we can solve your storage throughput challenges.

About Terascale

Terascale develops high throughput, high capacity and cost-effective storage solutions. Its unique storage appliance approach is changing the dynamics of the performance driven computing market, enabling existing users to do more for less while enabling new users to maximize the capabilities of their processing infrastructure. Founded in 2005, Terascale is based in Avon, MA.



Terascale, Inc.
145 Bodwell St.
Avon, MA 02322
Tel: 508-588-1501
Email: sales@terascalacom
www.terascalacom