

# The Economics of High-Throughput Sequence Analysis

---

[www.logicaldepth.com](http://www.logicaldepth.com)  
[econ@logicaldepth.com](mailto:econ@logicaldepth.com)

Summer 2004

This paper was written to help the planning and selection of bioinformatics infrastructure for high-throughput sequence analysis. An overview of the more common options for increasing sequence analysis capacity is presented along with heuristics for choosing between specialized acceleration hardware and commodity computing clusters.

The logo for Logical Depth, featuring the words "Logical Depth" in a grey, sans-serif font centered within a white rectangular box.

Logical Depth

Copyright ©2004 Logical Depth, Inc. All Rights Reserved.

Logical Depth, Inc. logos, and trademarks or registered trademarks of Logical Depth, Inc. or its subsidiaries in the United States and other countries.  
Copyright ©2004 Logical Depth, Inc. All Rights Reserved.

Other names and brands may be claimed as the property of others. Information regarding third party products is provided solely for educational purposes. Logical Depth is not responsible for the performance or support of third party products and does not make any representations or warranties whatsoever regarding quality, reliability, functionality, or compatibility of these devices or products.

## Summary

The growth rate of the biological sequence data that needs to be analyzed is outpacing the standard rate of improvement in underlying computing infrastructure. For higher throughput sequence analysis, the choices are commonly special-purpose dedicated hardware or commodity cluster computing. We will explore the pros and cons of each option and outline a general framework for evaluating alternatives.

## Accelerating Sequence Analysis

### *Commodity Clusters*

IBM, HP, Dell, Sun and various more specialized cluster hardware vendors such as Appro, CDC, Linux Networx, Microway, RLX, etc. sell commodity computing clusters to bioinformatics departments. Some organizations even build their own computing nodes from commodity components. The most common examples are Linux clusters running on x86 CPUs from Intel or AMD.

Cluster Advantages:

- Flexibility: General purpose hardware runs in-house and 3<sup>rd</sup> party apps
- Scalable: Add nodes as needed
- Integration: Same applications, more nodes
- Upgrades: New nodes improve with commodity hardware advances

Cluster Disadvantages:

- Power Consumption: electricity and cooling costs
- Management: cluster management, system administration

### *Special-Purpose Dedicated Hardware*

Time Logic and Paracel are representative vendors selling special-purpose hardware to accelerate bioinformatics.

Special-Purpose Hardware Advantages:

- Power Consumption: use less power than comparable cluster
- Management: fewer computing nodes than comparable cluster

Special-Purpose Hardware Disadvantages:

- Special-Purpose: Only run vendor-provided apps
- Coarse Scalability: Scalability in costlier steps than per-node scaling
- Upgrades: Upgrades lag commodity hardware, at Vendor's discretion
- Integration: Change workflow to integrate special-purpose hardware

## **Comparing Clusters and Special-Purpose Dedicated Hardware**

The qualitative differences between clusters and black boxes should be taken into account when comparing price/throughput. A flexible commodity computing infrastructure can more readily adapt to shifts in algorithms or applications, or a mix of commonly used and in-house applications. When special-purpose hardware isn't being used for the applications the vendor provided, it can't be used for anything else. In situations where there isn't the physical space or power infrastructure required for a computing cluster, and locating the computing resource elsewhere is not an option, then special-purpose hardware might be a better fit.

Ultimately, a decision for one or the other should be based on rational economic analysis. Calculating the price/throughput ratios for the sort of sequence analysis your organization does is essential. Taking into account the costs of what are more qualitative factors can be difficult, but even a rough estimate is worth the effort.

### **Price/Throughput Example Calculations**

Let's say  $B$  is the basic price/throughput for special-purpose ("black box") hardware

$$B = (\$B_{hardware} + \$B_{support}) / B_{throughput}$$

We can refine this price/throughput estimate to account for the special-purpose nature of the hardware, since if it isn't running a vendor supplied application it's not running anything:

$$B = utilization * (\$B_{hardware} + \$B_{support}) / B_{throughput}$$

Let's say  $C$  is the basic price/throughput for a commodity cluster:

$$C = (\$C_{hardware} + \$C_{support}) / C_{throughput}$$

Accounting for electricity and management costs of a cluster:

$$C = (\$C_{hardware} + \$C_{support} + \$electricity + \$management) / C_{throughput}$$

We'll work through an example scenario where about \$200,000 is spent on cluster or special-purpose hardware and calculate example price/throughput for each with example numbers for throughput, etc. Use evaluation numbers in the real world.

Example specialized hardware price/throughput calculation:

$$B = \$200,000 / (100M \text{ sequence searches} / \text{hour}) = \$2000 / M\text{search} / \text{hour}$$

Suppose however that the specialized hardware is only really utilized half as much as cluster hardware because the equivalent cluster hardware can run a wider variety of applications. Then *utilization* = 50% and:

$$B = \$2000 / M\text{search} / \text{hour} * \text{utilization} = \$4000 / M\text{search} / \text{hour}$$

Example cluster hardware price/throughput calculation, assuming \$power + \$management = \$100,000

$$C = ( \$200,000 + \$100,000 ) / (100M \text{ seq searches} / \text{hour}) = \$3000 / M\text{search} / \text{hour}$$

The specifics will vary per application, cluster, black box solution, utilization, etc. In plain English, this particular example illustrates a case where the flexibility of clustered computing for a particular set of applications outweighs the power and management costs.

## Comparison Problems

Ultimately a clear price/throughput comparison taking into account all factors is the best way to make an infrastructure decision. After studying a number of problematic comparisons over the years we hope to highlight some of the more common pitfalls in order to make hidden assumptions clearer.

## Time Machine Comparison

Some black box vendor comparisons, perhaps written up some years ago but still available online, give the impression that a black box offering is the equivalent of more than, say, 1000 CPUs. On closer reading the CPUs used in this comparison might be 1GHz Pentium 3 CPUs. This highlights the relentless progress of Intel and AMD commodity hardware, since a cluster vendor could just as well pitch a modern 250 CPU cluster as the equivalent of 1000 CPUs from a few years ago.

## Two-Budget Comparison

Clusters do have higher management and electricity costs than most special-purpose hardware options, but even fairly accounting for these differences doesn't close the gap when, for example, a \$200,000 special-purpose hardware solution is compared to a \$50,000 linux cluster. When it comes to price/throughput, the questions and comparisons should assume the budget to solve a problem with clusters is about the same as that to solve the problem with special-purpose hardware. If anything, cluster budgets might be higher than special-purpose hardware budgets since a broader user base, or more departments, can use a flexible computing resource.

## Blank Slate Assumption

What if there's already special-purpose hardware? One argument of special-purpose hardware vendors is that applications can be shifted between clusters and specialized hardware to benefit from the comparative advantages of each option. For instance if an organization is running a cluster to do BLAST or HMMER as well as in-house application, the BLAST and HMMER jobs could be allocated to the special-purpose hardware and the in-house code can run on the cluster since those won't be available to run on special-purpose hardware. By the same token, groups that already have some special-purpose hardware but are looking for increased throughput should still consider commodity clusters for additional throughput if it makes economic sense. The best solution in many of these cases might be to let the older special-purpose hardware run some or all of the jobs that best suit it, and to use new commodity clusters for additional capacity.

What if there's already a cluster? Groups in the market for higher throughput often already have a cluster in place that they're considering upgrading or replacing. In these cases the management and facilities costs for a cluster have already been assumed and aren't necessarily going to disappear with the purchase of special-purpose hardware. Also the management overhead of additional nodes might be much lower if there's already a large cluster in place. For certain applications there might also be optimized software available for higher throughput on the existing cluster, which can change the equation significantly.

## **Faster Sequence Analysis Software**

Using commodity clusters it's possible to see greater throughput with optimized software, an option that isn't available with special-purpose hardware, where the software is whatever the vendor provides. The open nature of commodity hardware means that even for a particular application there can be several different implementations that differ by ease of use or performance. Buying optimized software for a commodity clusters can offer much better price/performance than buying specialized hardware, while retaining the flexibility of commodity clusters.

Let's consider an accelerated software application, *S*, that's 8x faster than the commodity software option and costs \$100,000, compared with our earlier calculation of commodity hardware with commodity software:

$$C = \$200,000 + \$100,000 / 100M \text{ seq searches / hour} = \$3000 / M \text{ search / hr}$$

$$S = \$100,000$$

$$C + S = \$400,000 / 800M \text{ seq searches / hour} = \$500 / M \text{ search / hr}$$

In this example it turns out that optimized software on a commodity hardware cluster delivers better price/throughput than specialized hardware. The numbers we're using here are just examples, and we suggest any organization in the market for better price/throughput should do their own evaluations and calculations with real numbers and searches for the applications most relevant to them.

## **Conclusion**

A fair-minded economic analysis of special-purpose hardware, commodity hardware, and accelerated software for high-throughput applications of interest should result in meaningful price/throughput numbers on which to base a purchasing decision.

## Logical Depth

Logical Depth specializes in accelerating bioinformatics applications and algorithms for commodity CPUs. If your applications of interest include HMMER or Smith-Waterman, please consider us in your evaluations.

We are also interested in hearing about other algorithms or applications you would like to have accelerated for your commodity cluster.

## More Information

For the latest information about our product and services, please visit our website.

**Logical Depth:** <http://www.logicaldepth.com>

### Contact Us

Address: 650 Castro St. Suite 120-444

Mountain View, CA 94041

USA

Email: [econ@logicaldepth.com](mailto:econ@logicaldepth.com)

Tel: 866-682-9206

Copyright ©2004 Logical Depth, Inc. All Rights Reserved.

Logical Depth, Inc. logos, and trademarks or registered trademarks of Logical Depth, Inc. or its subsidiaries in the United States and other countries.  
Copyright ©2004 Logical Depth, Inc. All Rights Reserved.

Other names and brands may be claimed as the property of others. Information regarding third party products is provided solely for educational purposes.  
Logical Depth is not responsible for the performance or support of third party products and does not make any representations or warranties whatsoever regarding quality, reliability, functionality, or compatibility of these devices or products