Bloor

InDetail

InDetail Paper by Bloor
Author **Philip Howard**
Publish date **December 2015**

Blazegraph GPU

> **SYSTAP has implemented graphical processing units (GPUs) as accelerators for graph analytics... this approach represents approximately an order of magnitude improvement in cost-effectiveness.**

Author **Philip Howard**

Executive summary

Graph databases have significant advantages compared to other types of database when it comes to exploring and traversing data on the basis of relationships between data elements, particularly when those relationships are complex and varied. However, for graph analytics (as opposed to operational use cases) performance problems start to occur when graphs become very large (billions of nodes or more), especially where you are dealing with unknown relationships. The reason for this is because graphs have a property (or really an anti-property) known as "non-locality on disk". What this means is that where you do not know about relationships in advance you do not know where data will be stored and you cannot partition the database in any way that helps you. Since graph databases typically operate in-memory this in turn means that you spend a lot out of time paging data into and out of memory, which leads not only to what is known as a "cache thrash" but also to excessive use of the network and, since the network has a relatively high latency, this means that performance deteriorates.

Prior to the release of Blazegraph GPU there have been two approaches to overcoming this performance bottleneck. One is to use an architecture known as Bulk Synchronous Parallel. This reduces the traffic moved across the network. However, while it certainly reduces network traffic the benefits are relative rather than overwhelming. The second approach is to put the entire database in memory. This overcomes the performance and scalability problems associated with non-locality but does so at the expensive of very costly hardware: hundreds of terabytes of memory (and the additional hardware needed to sustain that) does not come cheaply. What SYSTAP, the company behind the development of Blazegraph, has done is to implement graphical processing units (GPUs) as accelerators for graph analytics. The company's figures, based on industry standard benchmarks, suggest that this approach represents approximately an order of magnitude improvement in cost-effectiveness compared to an all-in-memory approach, and that adding GPUs to the Blazegraph database (without GPUs) improves performance by better than two orders of magnitude.

Fast facts

SYSTAP LLC is the developer of the Blazegraph graph database, which has been under continuous development since 2006. Unlike the majority of vendors in this market, which tend to target operational and hybrid operational/query environments, SYSTAP is squarely focused on graph analytics and query. In addition to Blazegraph, the company market Blazegraph GPU, which is an add-on to Blazegraph enabling graph analytics to be accelerated using NVIDIA graph processing units. The company has also announced though not, at the time of writing released (it is due in Q1 2016), a product called DASL (pronounced "dazzle") that supports the development of analytic and statistical algorithms that will specifically run using GPUs. All three of these elements: the database, the extension to GPUs, and DASL, are discussed in this paper. However, the emphasis is placed on Blazegraph GPU and the DASL language that will accompany it, so we will be relatively brief in our description of Blazegraph. It is assumed that the reader has at least some familiarity with graph databases and their use cases, which include recommendation engines, cyber and other forms of security (including fraud detection), community detection and clustering, drug discovery and genomics, and fault prediction in industrial and IoT (Internet of Things) environments, amongst others.

Unlike the majority of vendors in this market, which tend to target operational and hybrid operational/query environments, SYSTAP is squarely focused on graph analytics and query.

> **...prior to the introduction of Blazegraph GPU there was no affordable product on the market that could analyse very large graphs with appropriate performance. From this perspective Blazegraph GPU is in a class of one.**

Key findings

In the opinion of Bloor Research, the following represent the key facts of which prospective users should be aware:

- Blazegraph is an (extended) RDF graph database also supporting property graphs.
- It supports SPARQL 1.1, the Tinkerpop/Blueprints and Sesame APIs and a graph mining API. You can also use the Gremlin graph traversal language and the product supports OWL (web ontology language).
- The product supports the development of domain specific languages whose syntax is converted into SPARQL queries at run-time.
- High availability, failover and load balancing are built in for the product's cluster-based solution. Single server and embedded editions are also available.
- Existing Blazegraph implementations, as well as new deployments, may be extended by the use of NVIDIA GPUs. Currently support is provided for hundreds of GPUs, enabling exploration of hundreds of billions of edges. There are plans to expand these to thousands of GPUs supporting trillions of edges.
- SPARQL queries are translated into suitable code for the GPUs by the software so there will be little or no change required when upgrading to a GPU-based environment.
- SYSTAP will be introducing a programming environment called DASL early in 2016 to support the development of graph algorithms within the Apache Spark ecosystem specifically optimised for GPUs. Specific algorithms will be developed by SYSTAP and will be made available according to customer demand.
- Blazegraph and Blazegraph GPU are specifically targeted at large scale, complex graph analytic environments, especially where relationships are unknown in advance.
- SYSTAP will be announcing a hardware appliance available in Q1 2016, which will bundle Blazegraph with GPU hardware in addition to cloud-based Blazegraph GPU offerings.

The bottom line

There are relatively few companies that specifically address graph analytics as opposed to analytics within the context of operational applications, which tend to be much simpler and at smaller scale. Of those that do address these sorts of problems some (for example, Teradata and Oracle) are limited to environments where relationships are known in advance. In other words, these products are not suitable for discovery, data mining (or science) and other complex analytic problems. Of the remainder, these are either extremely expensive and out of the practical reach of most organisations or, in some instances, vendors have attempted to build low cost solutions but these have failed the test of performance and scale. To summarise, prior to the introduction of Blazegraph GPU there was no affordable product on the market that could analyse very large graphs with appropriate performance. From this perspective Blazegraph GPU is in a class of one.

Blazegraph

Blazegraph is an open source graph database available via either a GPL or a commercial license. The current release is version 1.6. Unlike some other graph products, Blazegraph is targeted specifically at complex analytic environments rather than at operational systems. The product has been written entirely in Java (with a memory manager to manage the JVM heap) so will run on any platform that supports that language and it can be deployed either on a stand-alone server or on a cluster-based architecture and there is also an option called NanoSparqlServer that can be deployed as an embedded capability with access via a RESTful API. We will not be discussing this option further. The product ships with a fast, parallel loader.

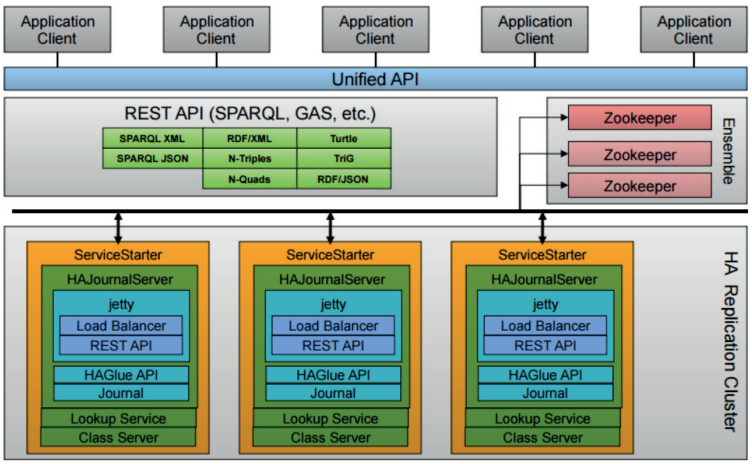
Stand-alone implementations have a limit of around fifty billion triples or quads where a quad is an extended triple whereby the context (name of the graph) is also stored. In theory there is no limit to the number of triples (or quads) that you can store if using a cluster-based environment. Clustered implementations include both high availability (using ZooKeeper – see **Figure 1** – which means that the cluster is self-healing as long as there is a quorum) and load balancing as well as cluster management and monitoring capabilities. For cloud-based deployments (private or public) multi-tenancy is supported. Public clouds supported include Amazon, IBM SoftLayer, and Cirrascale Cloud and we expect other cloud platforms to be announced in due course. Note that IBM SoftLayer and Cirrascale Cloud, in particular, support GPUs (see later). The company has indicated that there are plans for a hardware appliance available in Q1 2016, which will bundle Blazegraph with GPU hardware.

Fundamentally, Blazegraph is an RDF (resource description framework) database. In addition, Blazegraph also supports RDR (reification done right), which is an extension to the RDF standard, which allows you to add properties to the edges of a graph. In other words, RDR effectively means that you have the characteristics of a property graph while remaining, formally, an RDF database. Note that being able to add properties to edges significantly reduces node inflation in the graph and thereby improves both performance (fewer nodes to traverse) and scalability.

As an RDF database, Blazegraph supports the SPARQL 1.1 family of specifications including the query, update, basic federated query and service description components of that specification. In addition, SYSTAP supports the Sesame API as well as the Apache Tinkerpop project, where the latter includes both the Blueprints API and the Gremlin graph traversal language. Note that Gremlin is procedural whereas SPARQL is declarative. There is also an RDF graph mining API (GAS in **Figure 1**) which provides extensible graph mining that supports PageRank, Connected Components, Single Source Shortest Path, Breadth First Path and other graph algorithms.

In addition to RDF, SYSTAP also supports RDFS (RDF Schema), OWL (Web Ontology Language) and RDFS+ (RDFS with a bit of OWL). All of these are typically defined in relevant engines as a set of pre-defined inference rules. It is also worth mentioning that a number of SYSTAP's clients (notably in defence and life sciences) have developed their own domain-specific languages from the basic support offered in Blazegraph. These languages are converted into SPARQL queries at run-time.

Figure 1: Example of Zookeeper clustered implementations

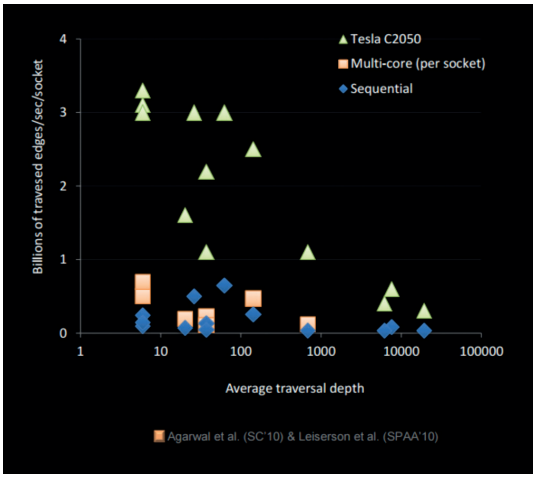


> **Unlike some other graph products, Blazegraph is targeted specifically at complex analytic environments rather than at operational systems.**

Finally, with respect to performance, pipeline parallelism is supported and there is a run-time query optimiser included within the database to ensure optimal performance for SPARQL queries. Note that an optimiser is especially important in graph databases where non-selective queries (that is, queries that cannot leverage an index – Blazegraph provides a mechanism for custom indices) are used. Complex graph problems often have this property. It is also worth commenting that the optimiser is capable of distinguishing between low latency queries (where the optimiser itself may be an obstacle to performance) and longer running queries. The optimiser is aware of resources such as memory utilisation and disk I/O operations.

> **SYSTAP is one of the few companies in the graph space to focus specifically on large-scale graph queries.**

Figure 2: Performance for Breadth First Search (BFS) algorithm



Blazegraph GPU

As noted, SYSTAP is one of the few companies in the graph space to focus specifically on large-scale graph queries. While these are not especially expensive, either in cost or performance terms, if you already know what relationships exist across your data, they become very expensive indeed if you are trying to discover relationships that you do not know exist. In effect, this precludes the use of conventional partitioning or sharding of the data (because you don't know how the data is organised) and you cannot, for the same reason, index the data in a suitable fashion. Historically, the only way to guarantee satisfactory performance is to load all of your graph data into memory. But this is horrendously expensive. In order to square this circle SYSTAP has partnered with NVIDIA to perform graph processing on that company's GPUs (graph processing units). The way that this works is that SPARQL queries are translated by the software to run on the GPU. This is done automatically so that users will not need to know how to program GPUs and, in the vast majority of cases, no change will be required to any existing SPARQL queries. Existing APIs for RDF and property graphs are fully supported.

The GPU option is available, commercially licensed, either as an add-on to an existing Blazegraph implementation or as part of a new deployment. At present, the software supports up to 256 GPUs, which should be adequate for graphs with edges measured at around one hundred billion. In the future the company intends to support thousands of GPUs in order to manage graph with trillions of edges.

In the opening section (the Executive Summary) of this paper we stated that this approach is approximately an order of magnitude more cost-effective than an all in memory approach. The next release of NVIDIA processors (Pascal) is expected to improve performance by another four times. We also stated that adding GPUs to an existing Blazegraph implementation was likely to improve performance by two orders of magnitude. We need to explore these claims in a little more detail. However, we will take it as read that GPUs have high performance characteristics in general and we do not intend to go into detail about the features of GPUs that enable this.

Figure 2 is taken from a presentation given at the GPU Technology Conference based on work done by Merrill, Garland and Grimshaw at the University of Virginia. It shows performance for the Breadth First Search (BFS) graph algorithm. This is interesting on two counts. Firstly, because it demonstrates that using GPUs (Tesla C2050 in this example) is faster than conventional techniques, regardless of traversal depth, ranging upwards from ten times better performance. Secondly, prior to this study it had been assumed that BFS was an

algorithm that was not well suited to the use of GPUs. In other words, this is a worst case example rather than the reverse, which makes these results even more significant than might otherwise be the case.

The other point we should reference is SYSTAP's own benchmarks comparing Blazegraph with and without GPUs. The example it quotes is one of the Lehigh University Benchmarks (LUBM) query number 9. Using Blazegraph on the LUBM 1000 University benchmark this query returns 172,632 results in just under 54 seconds. Adding GPUs reduces the response time to 187 milliseconds. That's an improvement of 290 times. Now, no doubt SYSTAP has chosen the most impressive of the LUBM tests to quote but, nevertheless, these are impressive figures.

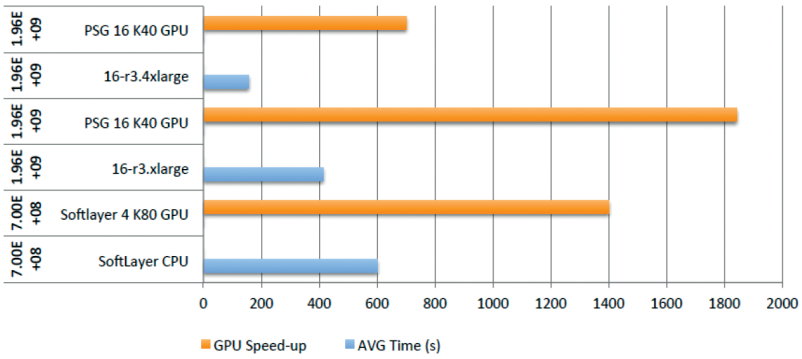
One further set of results is worth illustrating, as in **Figure 3**. However, this will take some explanation. The blue bars represent how much time was taken to run the BFS algorithm on various platforms of CPU-based Spark configurations and GPU-clusters. Thus, running on IBM's SoftLayer CPU configuration this analysis took just over ten minutes (601.34 seconds). The orange bar above it indicates how much faster it was when SoftLayer was implemented using Blazegraph GPU as opposed to Blazegraph on its own. Here the acceleration was 1400 times. In actual elapsed time that is 0.4284 seconds, which is why it is not times themselves that are shown on this chart.

Finally, it is worth commenting on price/performance. SYSTAP claims that the cost of Blazegraph GPU is around $16,000 per GTEP, where a GTEP represents a billion edges traversed per second. Comparable figures for in-memory processing at the same speed would be well into six figures.

DASL

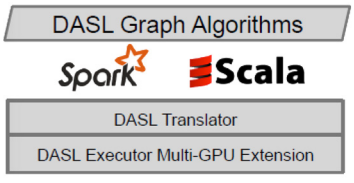
At present Blazegraph can only be extended through the use of NVIDIA GPUs. In principle there is no reason why other vendors' GPUs should not be used but suppliers in this market tend to use their own, proprietary, development environments. In the case of NVIDIA this is CUDA. This stands for Compute Unified Device Architecture, and it is a parallel computing platform and application programming interface (API) model that allows software developers to use a CUDA-enabled graphics processing unit for general purpose processing. The CUDA platform is a software layer that gives direct access to the GPU's virtual instruction set and parallel computational elements. However, despite the fact that you can program in C, C++ or Fortran, parallel programming is hard. As result, SYSTAP has developed DASL to make this process easier using its patent-pending technology. Its components are shown in **Figure 4**.

Figure 3: BFS comparative figures for GPUs



DASL is a functional, domain specific language for developing graph and machine learning algorithms. Potential applications include not just graph algorithms but also recommendation systems using collaborative filtering, neural network techniques, clustering algorithms, and a number of others. Over time SYSTAP will develop a number of these – based on customer demand – to be delivered out of the box. In the meantime, DASL will allow users to develop their own algorithms. As can be seen DASL leverages both Apache Spark and Scala and then the relevant code is automatically converted into CUDA so that developers do not have to worry about understanding parallel programming.

Figure 4: Components of DASL



The vendor

SYSTAP is a privately owned company, which was previously known as Scalable Graph Technologies. It is based in Washington, DC and first started developing what is now Blazegraph in 2006, based on research from the University of Utah. The company is privately funded and has received research and development funding from DARPA to develop its GPU technologies. This is significant because it gives SYSTAP an introduction to the defense and government markets in the United States. Originally, Blazegraph was called BigData. As readers will be aware big data as a concept has taken on a life of its own since then, and in 2014 the company decided to change the name of its product, since – through no fault of SYSTAP's – it was confusing. Readers may also see mention of MapGraph, which was the provisional name for Blazegraph GPU before this was officially announced.

The company has a number of significant users including DARPA, Yahoo!, The British Museum, Harvard Medical School, Autodesk, EMC2, and Wikimedia. The last of these is particularly interesting because it has made its evaluation criteria and its assessment of various graph database products available for public consumption. The company also has a number of partners apart from NVIDIA, which gives the company some presence outside North America, notably in Germany and Russia.

“The company has a number of significant users including DARPA, Yahoo!, The British Museum, Harvard Medical School, Autodesk, EMC2, and Wikimedia.”

Summary

A number of companies have been toying with the use of GPUs for some time. For example, several years ago we saw a demonstration of GPUs being used to speed up SQL processing. However, while beneficial, the cost-effectiveness of these approaches has been limited. The marriage of graph processing with graphical processing, on the other hand, looks to be very significant indeed. Moreover, when you consider that independent forecasts suggest that a quarter of all large enterprises will invest in graph technology by 2017 then not only do graph databases have a significant future, but SYSTAP looks to be very well placed to capitalise on the growth of interest in this technology.

FURTHER INFORMATION
Further information is available from
www.BloorResearch.com/update/2272

> **SYSTAP looks to be very well placed to capitalise on the growth of interest in this technology.**

About the author

PHILIP HOWARD
Research Director / Information Management

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.

After a quarter of a century of not being his own boss Philip set up his own company in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director focused on Data Management.

Data management refers to the management, movement, governance and storage of data and involves diverse technologies that include (but are not limited to) databases and data warehousing, data integration (including ETL, data migration and data federation), data quality, master data management, metadata management and log and event management. Philip also tracks spreadsheet management and complex event processing.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to ***IT-Director.com*** and ***IT-Analysis.com*** and was previously editor of both ***Application Development News*** and ***Operating System News*** on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and written a number of reports published by companies such as CMI and The Financial Times. Philip speaks regularly at conferences and other events throughout Europe and North America.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master), dining out and foreign travel.

Bloor overview

Bloor Research is one of Europe's leading IT research, analysis and consultancy organisations, and in 2014 celebrated its 25th anniversary. We explain how to bring greater Agility to corporate IT systems through the effective governance, management and leverage of Information. We have built a reputation for 'telling the right story' with independent, intelligent, well-articulated communications content and publications on all aspects of the ICT industry. We believe the objective of telling the right story is to:

- Describe the technology in context to its business value and the other systems and processes it interacts with.
- Understand how new and innovative technologies fit in with existing ICT investments.
- Look at the whole market and explain all the solutions available and how they can be more effectively evaluated.
- Filter 'noise' and make it easier to find the additional information or news that supports both investment and implementation.
- Ensure all our content is available through the most appropriate channel.

Founded in 1989, we have spent 25 years distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services, events and consultancy projects. We are committed to turning our knowledge into business value for you.

Copyright and disclaimer

Bloor

2nd Floor
145–157 St John Street
LONDON EC1V 4PY
United Kingdom

Tel: **+44 (0)20 7043 9750**
Web: **www.Bloor.eu**
email: **info@Bloor.eu**