



The GPU Accelerated Database

Eli Glaser
Senior Software Engineer
eglaser@gpudb.com

801 N Quincy St
Suite 601
Arlington, VA 22203

GPUdb - TL;DR



- In-memory distributed database using GPUs for processing
- Ultrafast ingest and analysis of billions of objects
- Built in visualization
- Full text search

GPUdb Overview



- A big data object store and calculation engine that is accelerated with NVIDIA Graphical Processing Units (GPUs)
- Enables big data analytics on the fly with streaming near real time data
- Calculate multi-dimensional algorithms with big data in sub-second time
- Native geospatial object support (points, shapes, tracks) for visualization as an image or video
- Full text search including wildcards
- High Performance Computing with commodity hardware costs
 - Scalable from a single laptop to a large cluster
 - Order of magnitude performance gain compared to CPU based clouds
 - Order of magnitude power reduction savings
 - Order of magnitude (or more) cost savings

GPUdb Features



- Abstracts distributed GPU processing from software developers
 - Memory management
 - Cluster wide GPU job scheduling
 - Automatic sharding and indexing
- Developers dynamically define data schemas
- Includes hardware accelerated geospatial, temporal, financial and machine learning processing functions
- Simple HTTP Rest-like API
 - Available API language wrappers: JavaScript, Java, Python, C++, C#
 - Trivial to add new language wrappers

GPUdb advantages in the NoSQL space



- Orders of magnitude faster than relational and 'NoSQL' competitors
 - Particularly for queries that need to scan all the data (i.e. count, sum)
- Reduced development costs for data scaling and data analytics
 - GPUdb does not require complicated key sharding techniques that some NoSQL players require (MongoDB, Hbase, Cassandra)
- Vastly smaller power and space footprint for greater computational capability

GPUdb Technical Challenges

- Memory Management
 - Disk->[CAPI]->RAM->vRAM
- Distributed GPU job coordination and scheduling
- Aligning computational cores with the data
- Performance, performance, performance

GPUdb Achievements



● US Army INSCOM

- In-memory computational engine for all data with geospatial and/or temporal components
- Integration with Apache Accumulo including per-object access control
- SGI UV2000 – 10TB of RAM and 16 K40 GPUs



● USPS

- In production ingesting and processing billions of objects
 - Geospatial breadcrumbs of USPS carriers
 - Mail delivery optimization
 - Multiple SGI UV2000s with 60+ Tesla K40s



● IDC HPC User Forum

- Won [IDC HPC Innovation Excellence](#) Award at SC14

GPUdb and OpenPOWER



- GPUdb is fully integrated and optimized on OpenPOWER hardware and software
 - IBM Power8
 - Ubuntu 14.04 - Little Endian
- NVIDIA Tesla K80 tested and certified
- IBM CAPI Large Scale Flash Memory Integration underway
- NVIDIA NVLink hardware beta testers

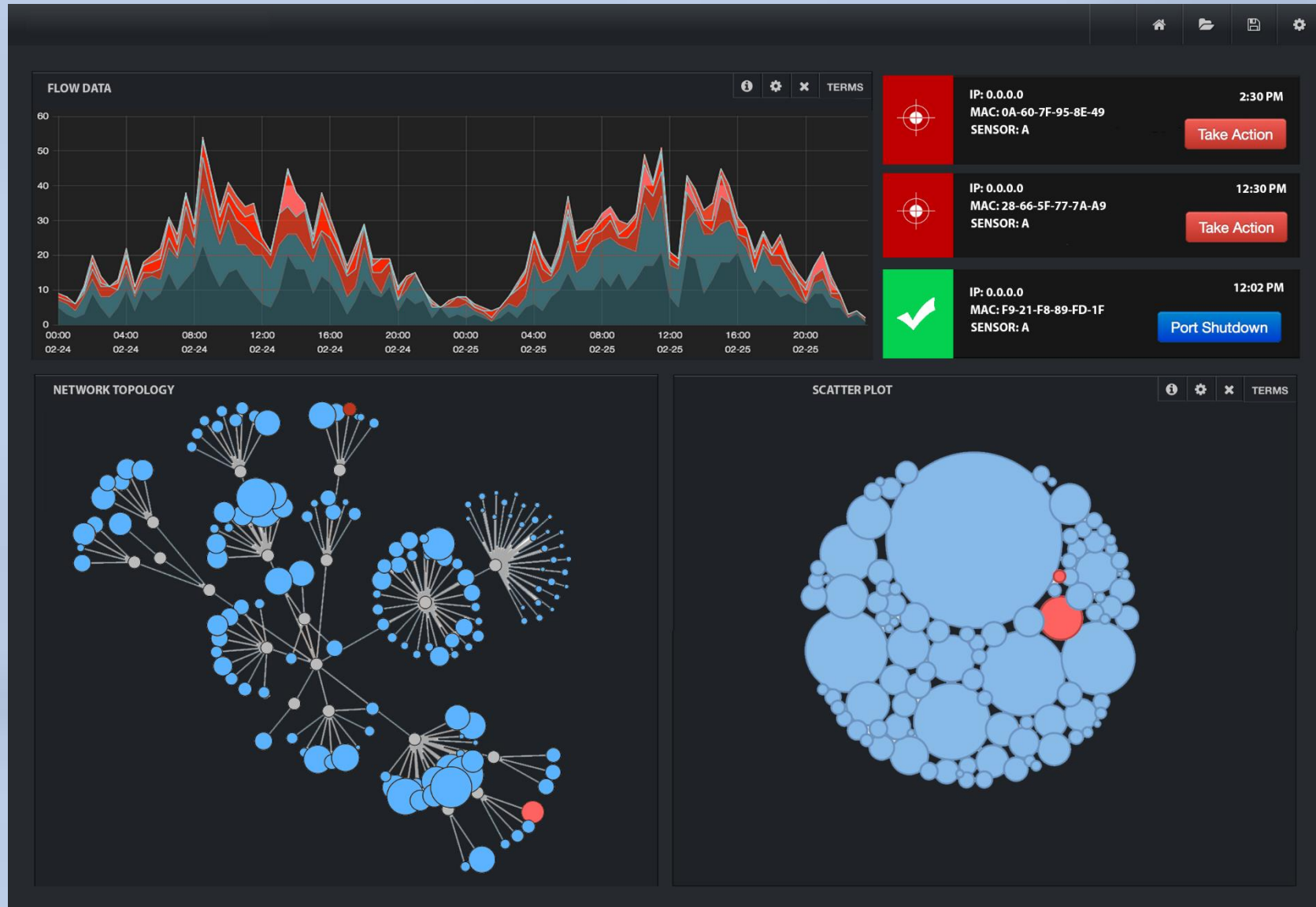
Come see us at the IBM booth

GPUdb and Cyber Intelligence



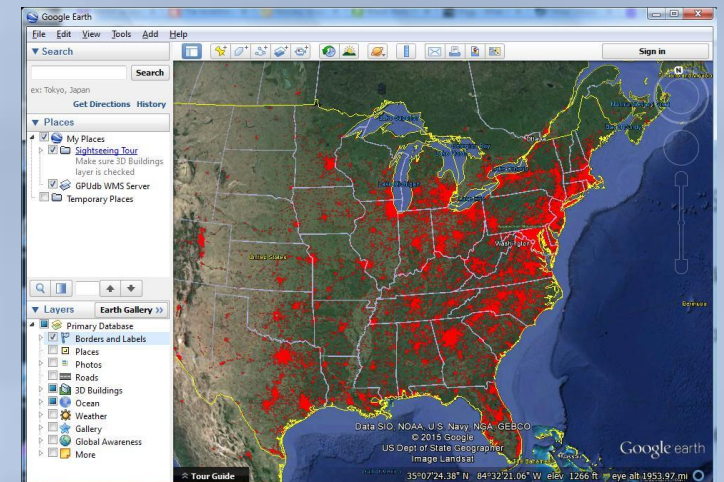
- GPUdb is capable of ingesting network 'flow' data at very high speeds
- Massive threading capability allows for computationally intensive deep packet processing analytics
- Native IPv4 and IPv6 attribute types for advanced network oriented query construction

GPUdb and Cyber Intelligence



GPUdb and GeoSpatial Processing

- Native understanding of geospatial objects including points, shapes, tracks
 - Shape processing: within, contains, intersection, etc
 - Convex hull
- Track analytics
- Includes a full embedded WMS server for easy integration with visual mapping frameworks
 - Google Earth / Google Maps
 - Cesium
 - OpenLayers
 - ESRI ArcGIS JS API



Real Time MGRS Clustering



Real Time Server-Side Video Generation

(Click Play)



The screenshot displays a web application interface for real-time video generation. The browser window shows the URL `https://localhost.teaminvertix.int/owf/#guid=5e697e86-b7d9-4943-9a32-66b79ce5f2c5`. The application is titled "OZONE Widget Framework" and includes a "Sample" tab and a "Notifications" area.

The main interface is divided into three primary sections:

- Infinity Demo:** Contains controls for uploading KML files and Gaia/WMS playback. The "Upload KML" section has a "Browse..." button and "Submit" and "Clear" buttons. The "Gaia/WMS Playback" section has radio buttons for "Disjoint", "Overlap", and "Grow", and a "Load Gaia/WMS Video" button.
- CPCE 3d:** A 3D map view showing a satellite image of a city area. It includes a "Layers" panel, "Tools", and "Create" options. A timeline overlay is visible at the top of the map, showing a range from 9:01 am to 9:02 am on 6/21/2011.
- Timeline HD:** A bar chart showing "counts" over time. The x-axis represents time from 17:53 to 18:02. The y-axis represents counts, ranging from 0 to 1,000. The chart shows several bars of varying heights, with the highest bar at 17:56. The chart is zoomed to the "1h" view and covers the period from 2011-06-21 to 2011-06-21.

The bottom of the screenshot shows a "Screencast-O-Matic.com" watermark and the application's taskbar.

Real Time Server-Side Heatmap Video Generation

(Click Play)

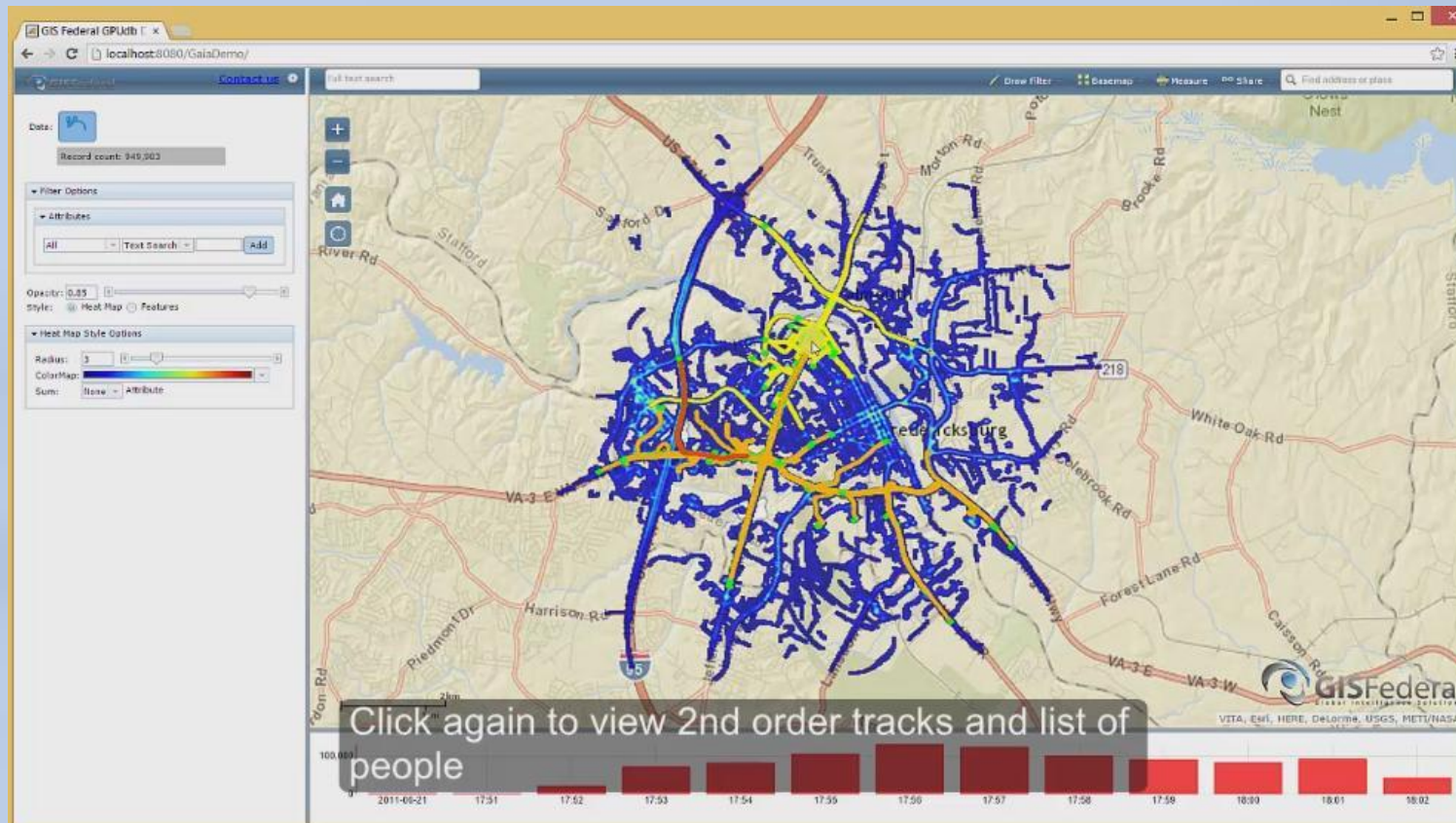


The screenshot displays a web application interface for OZONE Widget Framework. The browser window shows the URL `https://localhost.teaminvetix.int/owf/#guid=5e697e86-b7d9-4943-9a32-66b79ce5f2c5`. The main interface is divided into several sections:

- Browser:** Firefox browser window with the OZONE Widget Framework logo and navigation icons.
- Infinity Demo Sidebar:**
 - Filters:** Features, Tools
 - Upload KML:** Browse... (No file selected), Submit, Clear
 - Gaia/WMS Playback:** Disjoint, Overlap, Grow (selected), Load Gaia/WMS Video, Clear
- Map Area:** CPCE 3d map showing an aerial view of a residential area. A video player overlay shows a timeline from 6/21/2011 1:02:30 pm to 1:02 pm. The map includes a Google Earth logo and coordinates: 38°17'28.84" N, 77°28'06.83" W, elev 20 m, eye alt 1.09 km.
- Heatmap:** A bar chart showing the number of points over time. The x-axis represents time from 17:53 to 18:02. The y-axis represents the number of points, ranging from 0k to 2k. The bars show a significant increase in points starting around 17:54.
- Bottom:** Screencast-O-Matic.com watermark and window titles for CPCE 3d and Infinity Demo.

GPUdb and Track Analytics

- Use cell phone tracks to find when people might have been in contact with 'patient zero'



GPUdb Entry-Level Cluster Configuration



● 5 node Cluster

● Single Node \$1,008.00

● 1U used server from EBay – \$869.00

● 2x Intel Xeon X5650 (6-core, 2.66 GHz)

● 72 GB RAM

● 3 TB HDD

● 1x NVIDIA GTX 750Ti GPU - \$140.00

● 640 cores

● 2 GB vRAM

● Maxwell Architecture

Total Price: about \$5k

Able to query and render over 2 Billion Tweets in ~1 second

GPUdb Mid-Level Cluster Configuration



- 2 node Cluster

- Single Node

- 2U SuperMicro Server

- 2x Intel Xeon E5-2690 v3 (12-core, 2.60 GHz)

- 512 GB RAM

- 3 TB SSD

- 2x NVIDIA K80 GPU

- 2x2496 cores per card

- 2x12 GB vRAM per card

- Kepler Architecture

Total Price: about \$50k

Able to query and render 15+ Billion Tweets in ~1 second

GPUdb Useful Links



- [GPUdb Homepage](http://www.gpudb.com) – <http://www.gpudb.com>
- [GPUdb Demo Site](http://www.gpudb.com/gaiademo) – <http://www.gpudb.com/gaiademo>
- [GPUdb Tutorial video](https://www.youtube.com/watch?v=CNK7Mr5h8k0) - <https://www.youtube.com/watch?v=CNK7Mr5h8k0>
- [IDC HPC User Forum presentation](https://www.youtube.com/watch?v=fY6FUOsUZKY) - <https://www.youtube.com/watch?v=fY6FUOsUZKY>
- [IDC HPC Innovation Excellence Award](http://www.idc.com/getdoc.jsp?containerId=prUS25250214) - <http://www.idc.com/getdoc.jsp?containerId=prUS25250214>
- [Datanami GPU powered Terrorist Hunter Article](http://www.datanami.com/2014/10/08/gpu-powered-terrorist-hunter-eyes-commercial-big-data-role/) - <http://www.datanami.com/2014/10/08/gpu-powered-terrorist-hunter-eyes-commercial-big-data-role/>
- [SGI, NVIDIA, and GIS Federal INSCOM Article with UV2000 and 16 Tesla K40s](http://www.sgi.com/company_info/newsroom/press_releases/2014/april/gis_federal.html) - http://www.sgi.com/company_info/newsroom/press_releases/2014/april/gis_federal.html

We're Hiring!



- info@gpudb.com

Come see us at the IBM booth



The GPU Accelerated Database