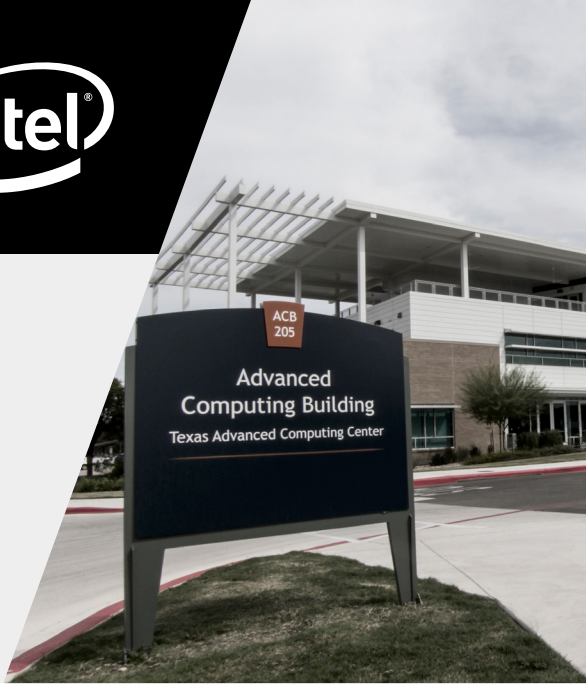
HPC BIG DATA ANALYTICS, A POTENT RESEARCH TOOL

When the Texas Advanced Computing Center launched Wrangler, a high-performance computing platform specifically designed for big data and analytics powered by Dell EMC and Intel®, it vastly expanded the world's research horizons



TACC
TEXAS ADVANCED COMPUTING CENTER

Higher education research facility | United States

Business needs

Traditional high-performance computing (HPC) is compute-intensive. But with research problems increasingly requiring analytics of larger data sets, including both structured and unstructured data — often in real time — HPC needs extreme infrastructure performance. That's why the Texas Advanced Computing Center sought to design and build an HPC platform optimized for both the computing and the storage demands of big data and analytics.

Solutions at a glance

- Big Data
  - Cloudera Hadoop
  - Data Analytics
- Cloud Solutions
- High-Performance Computing

Business results

- Brings innovation to market faster
- Increases cycles of learning
- Improves engineering efficiency
- Expands research options and inspires innovation with next-gen analytics

Accelerates analytics up to **20x**

Reduces I/O latencies by up to **99%**

High-performance computing (HPC) centers have grown in number worldwide, but they often cannot provide the storage performance (i.e., IOPS) that big data analytics needs. That's why the Texas Advanced Computing Center (TACC) at the University of Texas in Austin introduced "Wrangler," an HPC platform specifically designed for big data, data analytics and data-intensive applications.

"At TACC, our HPC clusters help address the computational problems that researchers can have across a range of research disciplines, and now with Wrangler's data-driven architecture, we have expanded that mission," says Niall Gaffney, TACC's director of data intensive computing. "Traditionally, HPC has served engineering fields that use physical equations to model the world, but more and more researchers need high-performance data analytics, which isn't what conventional HPC is optimized for."

Data-driven HPC opens new research horizons

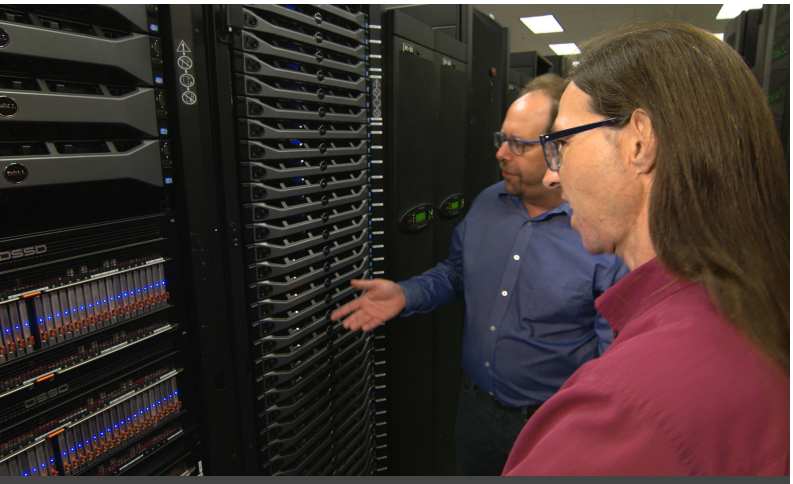
According to Gaffney, Wrangler's data-driven design enables researchers to explore both structured and unstructured data, static or dynamic, in ways they could not do before because of traditional HPC's I/O or storage limitations. "We designed Wrangler around storage, whether researchers need to tap into 10 petabytes of long-term storage to preserve their data or access half a petabyte of DSSD flash memory to compute against data without having the boundaries of standard disk based file systems getting in the way."

To process big data, perform data analytics and run data-intensive applications, Wrangler's three core HPC components are:

- **A compute platform** consisting of 96 Dell EMC PowerEdge R630 rack servers, powered by two 12-core Intel® Xeon® processors E5-2680 and with 128 GB of DDR4 volatile memory. Each analytics node features both Mellanox® InfiniBand® FDR and 40 Gb/s of Ethernet connectivity. Users can access up to 3,000 cores to compute against the multi-terabyte data sets they can store in Wrangler.
- **A half-petabyte rack-scale flash storage system** consisting of 10 Dell EMC DSSD D5 appliances. DSSD provides dense shared rack-scale server flash to up to 48 servers via NVM Express (NVMe), utilizing one of the largest and highest-performance PCIe fabrics ever built. Each 5-rack unit appliance can support throughputs to the analytics cluster at rates up to 100 GB/s transactions up to 10 million IOPS, and ultra-low latencies as low as 100 µs.
- **A 10-petabyte Lustre file system** using Dell Storage for HPC with Intel Enterprise Edition for Lustre solution for long-term storage. It is replicated between TACC and Indiana University in case of disaster recovery. Both systems are linked to Internet2 via a 100 Gbps connection, giving Wrangler a maximum potential network throughput of 200 Gbps for ingesting and accessing data.

Importantly, Wrangler includes Cloudera Hadoop running on top of the Dell EMC DSSD D5's, a leading big data analytics processing platform. "Wrangler is able to provide what many users are calling 'limitless I/O' because they can gain the full benefit of its incredible throughput and IOPS, given the DSSD Hadoop plug-in that's coupled with the Cloudera environment," says Gaffney.

> *"We've pushed HPC's boundaries with data analytics, so our users can push their research boundaries, all with Dell EMC and Intel technology fully behind us."*
>
> Niall Gaffney
> Director, Data Intensive Computing.
> Texas Advanced Computing Center



"Solutions can be cut from weeks to hours by reducing I/O latencies by as much as 99 percent, using Wrangler's Dell EMC hardware powered by Intel processors."

Niall Gaffney
Director, Data Intensive Computing,
Texas Advanced Computing Center

Extreme performance density enables new insights

Before Wrangler, researchers attempting big-data analytics on traditional compute-intensive HPC were taxing the file systems and affecting other people's work. "That doesn't happen with Wrangler," Gaffney says. "Its data-driven architecture helps expand how we can address researchers' needs today, whether they're using Hadoop or Spark-based models for big data analytics or using tools developed in Python, R or other development environments arising from more traditional files systems and clusters."

For example, one researcher who was using a traditional compute-intensive HPC cluster needed hundreds of nodes to get enough I/O for his analysis. "But with just 10 nodes on Wrangler, he accelerated data analytics by 20x, showing the power of the Dell EMC and Intel combination," Gaffney says, noting that the boost in data throughput was largely responsible for the improvement in performance. "Plus, he derived new analytical insights he couldn't get before."

Another researcher and now a big Wrangler fan is Chris Mattmann, chief architect of the Instrument and Science Data System section of NASA and the Jet Propulsion Laboratory. With Wrangler, he can complete a complex code analysis that traditional compute-intensive HPC could not complete at all. "Big data workloads are I/O constrained, but Wrangler removes that bottleneck with the Dell EMC DSSD D5 flash storage," he says.

Cutting I/O latencies by 99%

Wrangler again showed its data prowess in analyzing the full genomes of different species to understand how traits are inherited through genetic markers. "That's a huge I/O problem, comparing a simple series of DNA's four components repeatedly but looking for repeating patterns in extremely large files," Gaffney says. "Solutions can be cut from weeks to hours by reducing I/O latencies by as much as 99 percent, using Wrangler's Dell EMC hardware powered by Intel Xeon processors."

Today, Wrangler enables a range of disciplines to vastly expand their investigative horizons. "Now, researchers can ask questions about their data they couldn't before, with workloads accelerated by up to 1,000 percent by Dell EMC servers and DSSD rack-scale flash appliance," Gaffney says. "One researcher analyzed closing prices on all New York Stock Exchange listings going back 96 years and was able to test economic and market theories at scales that weren't practical before."

This increased engineering efficiency with accelerated cycles of learning delivers these kinds of new insights to keep Gaffney and TACC's 160 staff members inspired. He attributes much of Wrangler's success to the support and collaboration of Dell EMC and Intel. "With Wrangler, we've pushed HPC's boundaries with data analytics, so our users can push their research boundaries, all with Dell EMC and Intel technology fully behind us."

> *"Researchers can ask questions about their data they couldn't before, with workloads accelerated by up to 1,000 percent by Dell EMC servers and DSSD rack-scale flash appliance."*
>
> **Niall Gaffney**
> Director, Data Intensive Computing,
> Texas Advanced Computing Center

Intel Inside®. Powerful Productivity Outside.

Learn more about Dell EMC Big Data solutions

Contact a Dell EMC big data expert

View all customer stories at Dell.com/CustomerStories

Connect on social

DELL EMC

intel