A Medium-Scale Distributed System for Computer Science Research: Infrastructure for the Long Term

Bal, H.; Epema, D.; de Laat, C.; van Nieuwpoort, R.; Romein, J.; Seinstra, F.; Snoek, C.; Wijshoff, H.

DOI
10.1109/MC.2016.127

Publication date
2016

Document Version
Final published version

Published in
Computer

License
Article 25fa Dutch Copyright Act

RESEARCH FEATURE

A Medium-Scale Distributed System for Computer Science Research: Infrastructure for the Long Term

Henri Bal, VU University Amsterdam
Dick Epema, Delft University of Technology
Cees de Laat, University of Amsterdam
Rob van Nieuwpoort, Netherlands eScience Center
John Romein, ASTRON
Frank Seinstra, Netherlands eScience Center
Cees Snoek, University of Amsterdam
Harry Wijshoff, Leiden University

The Dutch Advanced School for Computing and Imaging has built five generations of a 200-node distributed system over nearly two decades while remaining aligned with the shifting computer science research agenda. The system has supported years of award-winning research, underlining the benefits of investing in a smaller-scale, tailored design.

Like any science that heavily uses experimentation, computer science depends on quality physical research infrastructures based on supercomputers and clouds through which researchers can access platforms for large-scale collaborative experiments. Although these parallel and distributed production systems have many benefits for computational work, such as rapid results when running complex simulations, they are ill-suited for research that needs detailed hardware and software control. The difficulties of using production systems to conduct controlled, reproducible, distributed experiments on multiple resources are well known.

PlanetLab, a global research network that supports the development of new network services, federates existing resources to enable large-scale experiments on thousands of nodes. However, many applications use the nodes simultaneously, making it hard to obtain reliable performance measures. Grid'5000 is a from-scratch large-scale distributed system dedicated to computer science,[1] but like

0018-9162/16/$33.00 © 2016 IEEE

TABLE 1. Sample research projects and their impact for four distributed Advanced School for Computing and Imaging (ASCI) supercomputer (DAS) generations.

Generation	Research project and focus	Impact
DAS-1	LFC: user-level network protocol	400+ citations
	Albatross: wide-area algorithms	Foundation for many subsequent projects
	MagPIe: wide-area collectives	500+ citations, influenced MPICH-G2
DAS-2	Awari: solving the Awari game	One of 250 mathematics milestones described in Clifford Pickover's *The Math Book*
	Ibis and Satin: distributed programming system	700+ citations, SCALE 2008 award, Euro-Par 2014 award
	JavaGAT: grid programming toolkit	60,000+ downloads; led to OGF SAGA standard
	Koala: co-allocating scheduler	CCGrid 2012 keynote presentation
	rTPL: tool to measure network performance	Internet2 Land Speed Record
	Tribler and Cyclon: peer-to-peer protocols	3,000+ citations, best paper award P2P 2006
DAS-3	StarPlane: reconfigurable optical network	Keynotes at NorduNet and Terena; StarPlane architecture applied in major research and education networks
	INDL: resource-description framework	Used for GENI and Fed4FIRE
	VL-e: virtual eScience lab	€20 million in funding
	Robot dog: object recognition on a grid	AAAI 2007 Most Visionary Research Award
DAS-4	Glasswing: MapReduce on many-core architecture	Best student paper nominee SC 2014
	COMMIT/: modeling GPU data transfers	Best paper nominee CCGrid 2014
	Squirrel: quick launching of virtual-machine images	SC 2013 and HPDC 2014 proceedings
	WebPIE: distributed reasoning	500+ citations, SCALE 2010 award
	BTWorld: large-scale time-based datasets	SCALE 2014 award

AAAI 2007: 22nd Conf. Artificial Intelligence; CCGrid: IEEE/ACM Int'l Symp. Cluster, Cloud and Grid Computing; Euro-Par 2014: 20th Int'l Conf. Parallel and Distributed Computing; Fed4FIRE: Federation for Future Internet Research and Experimentation; GENI: Global Environment for Networking Innovation; HPDC 2014: ACM Int'l Symp. High-Performance Parallel and Distributed Computing; INDL: Infrastructure and Network Description Language; JavaGAT: Java Grid Application Toolkit; LFC: Link-Level Flow Control Protocol; MPICH-G2: Globus-enabled MPI (message-passing interface) over Chameleon; OGF SAGA: Open Grid Forum's Simple API for Grid Applications; P2P 2006: 6th Ann. Int'l Conf. Peer-to-Peer Computing; rTPL: Remote Throughput, Ping and Load; SC: Int'l Conf. High Performance Computing, Networking, Storage and Analysis; SCALE: IEEE Int'l Scalable Computing Challenge; VL-e: Virtual Laboratory for e-Science

most large infrastructures, it suffers from fragmented funding, which translates to incremental updating that degrades cohesiveness over time.

These large-scale systems cannot keep pace with the inevitable changes in computer science research because each shift requires a costly infrastructure realignment. To avoid this problem, the Dutch Advanced School for Computing and Imaging (ASCI) launched the first generation of its supercomputer project (Distributed ASCI Supercomputer [DAS]) in 1997 and adopted the strategy of building medium-scale distributed systems and replacing them as researchers' needs change. Over the past two decades, five generations of DAS have consistently supported a shifting research agenda:

- DAS system generation 1 (DAS-1) in 1997, which supported wide-area computing with homogeneous hardware and software and a dedicated asynchronous transfer mode (ATM) network;
- DAS system generation 2 (DAS-2) in 2002, which supported grid computing with Globus middleware;
- DAS system generation 3 (DAS-3) in 2006, which supported optical grids with photonically switched 10-Gbps links between all sites;
- DAS system generation 4 (DAS-4) in 2010, which supported heterogeneous computing, cloud computing, and green IT through hardware virtualization, accelerators, and energy measurements; and
- DAS system generation 5 (DAS-5) in 2015, which supports heterogeneous computing and big data through a wide variety of accelerators, larger memories and disks, and software-defined networking (SDN).

Each generation's hardware is built from scratch, with a total budget of €1.5 million consisting of grants from the Netherlands Organization for Scientific Research (NWO) and matching funds from DAS project participants. DAS-5 comprises six clusters of 200 computational nodes spread across participant institutions and integrated through a wide-area network (WAN). Preceding generations had four or five clusters, about the same number of nodes, and the same budget.

All five DAS system generations have followed the same two architectural

TABLE 2. DAS system components across generations.

Architectural aspect	DAS-1 (1997)	DAS-2 (2002)	DAS-3 (2006)	DAS-4 (2010)	DAS-5 (2015)
Clusters	4	5	5	6	6
Cores	200	400	792	1,600	3,252
CPU	200-MHz Pentium Pro	Dual 1-GHz Pentium 3	Dual 2.2–2.6 GHz AMD Opteron	Dual quad-core Xeon E5620	Dual eight-core Xeon E5-2630 v3
Interconnect	Myrinet	Myrinet	Myrinet 10G	QDR Inifiniband	FDR Infiniband
Wide-area network	ATM	Internet	Optical path	Optical path	Optical path + SDN

ATM: asynchronous transfer mode; SDN: software-defined networking.

principles—centralized change coordination and the support of highly interactive experimentation—which the sidebar "Core Architectural Principles" describes in more detail. The generations differ only in how they support a particular research focus. Table 1 lists research projects in the first four generations (DAS-5 is relatively recent) and their support characteristics, along with corresponding research themes and research awards. Although other infrastructures exist for these research themes, none has provided a near-permanent cohesive basis for experimental computer science research. To ensure that the DAS system remains cohesive, each site has adopted the same changes, some of which have persisted through generations. An example is the consistent use of a dedicated optical interconnect to link participant sites starting with DAS-3 and extending to the current DAS-5.

These core principles have given DAS participant institutions familiar tools that are readily available to support the specific needs of computer science researchers. With these tools, the institutions have reaped the benefits of long-term research in strategic areas such as video retrieval and distributed computing. Despite its moderate size, the DAS system has been remarkably successful even in international competitions for which scale matters. DAS system users—100 to 150 scientists—have repeatedly won prestigious awards, including the IEEE CCGrid's (IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing's) SCALE (Scalable Computing Challenge) competition (three times) and the National Institute of Standards and Technology (NIST) Text Retrieval Conference Video Retrieval Evaluation (TRECVID) competition (six times), and have been published in top journals. These accolades underline the return on a long-term investment in an infrastructure that is easy to set up and maintain while consistently facilitating experimentation in the areas of current interest to computer science researchers.

DAS-1: WIDE-AREA COMPUTING

Table 2 gives an overview of DAS system components for each generation. The original idea was to design a completely homogeneous distributed system consisting of four clusters, each of which would be at one of the four original universities—the University of Amsterdam, VU University Amsterdam (VU), Leiden University, and Delft University of Technology—and interconnected in a dedicated 6-Mbps ATM network.

In DAS-1, all nodes used the same CPU (Intel PentiumPro), local network (Myrinet), and OS (initially BSDI Unix and later RedHat Linux). The only configuration difference was the larger VU cluster, which was due to the university's extra investment in all DAS generations and provided a basis for comparing the performance of an algorithm distributed over clusters versus the same algorithm run in a single cluster.

Accessing low-level software

DAS-1 confirmed our assumption that the DAS system users could experiment with low-level systems software. Researchers investigating user-level network protocols, for example, could directly access the network interface and thus eliminate the OS from the critical communication path. We implemented Link-Level Flow Control (LFC),[2] a (then new) network protocol for Myrinet that consisted of a user-level library and new firmware for the network interface card. With LFC, we confirmed that programmable network interfaces increase flexibility and reduce communication overhead. Researchers also used DAS-1 to investigate how multiple geographically distributed resources can be combined to solve computationally intensive problems. This study became the basis for the later development of real-world applications, such as climate modeling and astrophysics, on DAS-4.

Figure 1 shows the performance of several simple parallel algorithms on DAS-1. Some programs even ran slower on four clusters with 15 nodes each than on one cluster with 15 nodes because part of the communication went over wide-area links, which are orders of magnitude slower than the local network. Most algorithms, however, could be optimized for wide-area systems by implementing latency hiding, load balancing, and message aggregation, for example, to reduce communication overhead. The optimized programs generally ran faster on multiple clusters and often attained a performance close to that of a single large cluster with the same total number of nodes. Some fine-grained programs like retrograde analysis (RA) remained inefficient on wide-area systems, as expected.

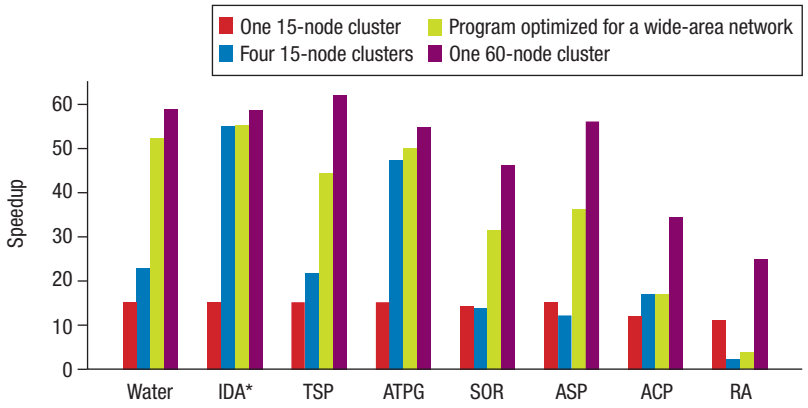

FIGURE 1. Performance of several parallel algorithms on Distributed ASCI Supercomputer 1 (DAS-1), the first generation of the Advanced School for Computing and Imaging's supercomputer project. All cluster and optimized performance results are relative to the original program run on one CPU. IDA*: iterative-deepening A*; TSP: traveling salesman problem; ATPG: automatic test-pattern generation; SOR: successive overrelaxation (method); ASP: all-pairs shortest paths (problem); ACP: arc consistency problem; RA: retrograde analysis.

Effects of varying bandwidth

To allow early experiments with different WAN speeds, VU researchers installed eight local ATM links in their DAS-1 cluster to create a virtual (experimental) system with the same hardware as the wide-area DAS-1 system and used delay loops to vary the ATM links' latency and bandwidth. Because the experimentation system used the same binaries and was identical in every way (except the ATM links) to the real wide-area system, researchers could use it to reliably analyze the sensitivity of parallel programs to wide-area latency and bandwidth.[3]

Early insights

Like the DAS system projects that would follow, early research projects required clean, laboratory-like experiments that were not easy to perform on production systems. Research on DAS-1 showed that as long as (often simple) wide-area optimizations were applied, distributed supercomputing was feasible for a much wider class of applications than previously believed. These insights are now common, and many were used to design programming systems for distributed supercomputing. MagPIe is an implementation of the message-passing interface (MPI) with operations optimized for hierarchical wide-area systems.[4] Ideas from MagPIe were later applied in MPICH-G2 (Globus-enabled MPI over Chameleon).

DAS-2: GRID COMPUTING

DAS-2 was based on the same principles as DAS-1, but it used the Globus middleware to enable grid experiments and the normal university network infrastructure with 1-Gbps Ethernet uplinks instead of the ATM interconnect. DAS-2 was used to study several grid programming environments and to conduct novel networking research.

Awari game solutions

The first major result obtained on DAS-2 was a solution to the Awari game, which found the best possible move among the 889 billion possible board configurations. The work was selected as one of Clifford A. Pickover's 250 mathematics milestones.[5]

Distributed programming

Another focus for DAS-2 was distributed programming with the Ibis and Satin projects. Ibis, which still exists, is a Java-centric programming system for high-performance applications on heterogeneous distributed systems.[6] Its core communication layer was designed for dynamically changing systems and it used the Java for Grid Application Toolkit (JavaGAT), a predecessor of the Open Grid Forum's (OGF's) Simple API for Grid Applications (SAGA) standard. As such, Ibis allows the transparent access of different resource types running a variety of grid middleware. Its SmartSockets component solved connectivity issues attributable to firewalls.

Subsequent research produced Satin, a Java-based programming system implemented with Ibis that could automatically run divide-and-conquer applications on distributed systems such as grids. Satin provided automatic cluster-aware load balancing, malleability, and fault tolerance. In concert with a programming environment suitable for many-core architectures, Satin became an important building block of the Cashmere system.

Multicluster scheduling

These experiments established a controlled testbed for understanding performance issues, and were the foundation for a move to more realistic platforms. Efforts such as combining DAS with the GridLab testbed and Grid'5000 showed that DAS system software could run outside the laboratory.

Research on the Koala multicluster scheduler began during this time. From its inception, Koala's key feature has been its support of processor coallocation for parallel applications through both scheduling policies and interfacing mechanisms to local cluster schedulers. When co-allocating a parallel application, either the user or the scheduler decides how to split it up into components for scheduling on a single site. Researchers later extended Koala to support the co-allocation of more application types, notably bags-of-tasks and workflows. This work paved the way for DAS-5 research on scheduling complete application frameworks, such as MapReduce.

Networking protocols

DAS was also used extensively to support networking research. DAS-1 used special testbed connectivity as well as the normal production network of SURFnet, which implements and maintains the National Research and Education Network of the Netherlands. This research model continued in future DAS system generations.

Many Internet protocols were defined in the 1980s, when both computer memory and wide-area bandwidth were scarce. DAS-2 supported experiments to understand and optimize the throughput of various high-speed transport protocols. Researchers wrote the Remote Throughput, Ping and Load (rTPL) package to measure network performance from an end-user viewpoint, which enabled experimentation with controlled mixes of many normal low-bandwidth Internet flows as well as tests with a few high-throughput flows. The latter experiments used extremely optimized new protocols that could destructively influence each other. rTPL research garnered the Internet2 Land Speed Record in both 2002 and 2004.

Peer-to-peer systems

DAS-2 research furthered the exploration of issues in peer-to-peer (P2P) systems, producing a gossip-based peer sampling service,[7] the Cyclon membership management protocol, and the Tribler social P2P system. Before designing Tribler, a P2P client based on BitTorrent,[8] researchers took extensive measurements of the worldwide BitTorrent P2P system to identify the main P2P system design challenges: decentralization, availability, and providing incentives. Performing these measurements required many IP addresses for contacting large numbers of peers. When designing and implementing Tribler, researchers emulated large numbers of Tribler peers on many DAS-2 nodes to test new components under controlled but realistic circumstances.

DAS-3: OPTICAL GRIDS

DAS-3 was built to accommodate the paradigm shift of hybrid networking,[9] in which optical photonic connectivity was introduced to augment Internet routing. Developments in photonics enabled wavelength-selective switches to dynamically route colors on StarPlane, a flexible dark-fiber infrastructure. DAS-3 was one of the first systems to use StarPlane along with several 10-Gbps static optical paths among SURFnet's sites. An additional 10-Gbps optical path from Amsterdam to Paris interconnected DAS-3 and Grid'5000.

Massive data aggregation

StarPlane enabled the study of more data-intensive applications. By implementing the Divine model checker on DAS-3, for example, researchers could experiment with consolidating the use of all cluster memories. This optimization was (and still is) important to model checking, a process that requires large memory because of state-space explosion. Combining multiple 10-Gbps links provided the required wide-area bandwidth, which allowed massive data aggregation—the combination of many small messages (state exchanges) into large asynchronous data transfers.

Infrastructure and Network Description Language

The hybrid networking and emerging heterogeneous computing infrastructures triggered another research problem: the need for an integrated information system that could find diverse resources. Typical existing solutions to resource location did not work across Internet layers and domains and did not cover the data-processing infrastructure.

Researchers proposed the Infrastructure and Network Description Language (INDL), a resource-description framework based on the Semantic Web, to define the topology of distributed infrastructures and locate resources across them. This approach eventually became the US National Science Foundation's Global Environment for Networking Innovation (GENI) testbed and the EU's Federation for Future Internet Research and Experimentation (Fed4FIRE) infrastructures.

Virtual Laboratory for eScience

One of the largest projects using DAS-3 was the Virtual Laboratory for eScience (VL-e), which created an e-science environment and performed research on workflow, distributed programming, resource management, and other methodologies. Funded in part by €20 million from the Dutch government, VL-e consisted of ASCI researchers who exploited Ibis, JavaGAT, Koala, INDL, and the VL-e workflow system, for example, to create a rapid prototyping e-science environment.

Robot dog movie analysis

One of the more creative results of DAS-3 research was an application that used a worldwide grid (which included DAS-3) to perform image recognition on movies captured by the Sony robot dog's webcam. The initial implementation used TCP to distribute the videoframes to clusters worldwide, with each cluster repeatedly processing a frame using a data-parallel MPI program. DAS-3

DAS GENERATIONS HAVE SHOWN THE BENEFIT OF LONG-TERM INVESTMENT IN INFRASTRUCTURE FOR COMPUTER SCIENCE RESEARCH'S SPECIFIC NEEDS.

was also used for performance-testing the application, which won the Most Visionary Research Award at the 22nd Conference on Artificial Intelligence (AAAI 2007). Researchers later used Java and Ibis to redesign the application, making it independent of platform and middleware, fault tolerant, and scalable. The Ibis version won the 2008 SCALE competition.

DAS-4: CLOUD COMPUTING AND GREEN IT

DAS-4 was designed as a testbed for experiments in cloud computing, heterogeneous computing, and green IT. Its core (CPUs, LAN, and OS) was still homogeneous, but ASCI added various accelerator types to the sites, which enabled comparisons between different GPU types within otherwise identical computational nodes. DAS-4 also contained power monitors, and its nodes could be set up with cloud middleware. Projects focused on a range of heterogeneous computing research, including projects centered on cloud computing and energy management.

Glasswing

Glasswing, a MapReduce framework on top of OpenCL, efficiently uses the resources of heterogeneous cluster environments by combining coarse-grained and fine-grained parallelism and aggressively overlapping computation, communication, memory transfers, and disk accesses. Researchers used Glasswing with DAS-4 to compare it against Hadoop on accelerator types. Glasswing yielded large performance improvements.

COMMIT/

An interesting research problem triggered by the parallel ocean program (which simulates ocean movement; www.cesm.ucar.edu/models/cesm1.0/pop2) was how to optimize data transfers between the host CPU and the GPU. The application has many small kernels that require many transfers; for each transfer, a choice must be made whether to use explicit copying, memory mapping, or Compute Unified Device Architecture (CUDA) streams. Within the COMMIT/ project, researchers developed a performance model that helps make this decision without having to try all possible combinations.

Squirrel and GreenClouds

DAS-4 was also used to study a variety of problems in cloud computing and green IT. The Squirrel project, for example, explored how to efficiently start a large number of virtual-machine images without creating the bottleneck that typically occurs in a high-performance cloud-computing environment.

In another project, researchers studied what it would take to virtualize networks and make them objects that compilers and advanced applications could program. In 2015, this research resulted in a collaboration with Royal Dutch Airlines and COMMIT/.

The green IT project GreenClouds aimed to decrease the energy consumption of high-performance computing systems. For example, the project studied the energy profiles of virtual machines and produced an energy-budget calculator and resource manager that supports the energy-efficient Ethernet (802.3a-z).

WebPIE and BTWorld

Another interesting line of research on DAS-4 is distributed reasoning. WebPIE is a system implemented on DAS with Ibis and Hadoop that can perform Web-scale reasoning by computing the so-called materialization (closure) of huge Resource Description Framework (RDF) graphs.[10] WebPIE won the SCALE challenge at CCGrid 2010 for solving a problem with 100 billion triples. In 2014, the DAS consortium won the SCALE challenge for the third time with a project (BTWorld) that performed large-scale analysis of time-based datasets, in particular monitoring data collected from BitTorrent servers over four years.

DAS-5: HANDLING DIVERSITY AND VOLUME

DAS-5, which became operational in May 2015, has not been in use long enough to report concrete research results from ongoing projects, but the new design includes 100-Gbps wide-area connectivity and the introduction of SDN. We also see three trends that will drive research efforts using DAS-5: heterogeneous computing, virtualization, and big data.

Heterogeneous computing

Heterogeneous computing on different types of accelerators is becoming increasingly important for supercomputing. Although DAS was initially designed as a completely homogeneous system, heterogeneity can be added as the research agenda dictates. DAS-5 is already incorporating different accelerators.

Virtualization

Distributed system components, such as computing, storage, networks, visualization, algorithms, and libraries, are rapidly becoming virtualized so that they can be easily grouped into services. Algorithms for data processing and simulations at levels from bare metal to platforms need to evolve to ensure scaling, fault tolerance, and energy efficiency. Data from

CORE ARCHITECTURAL PRINCIPLES

Two principles have enabled five generations of the Advanced School for Computing and Imaging's (ASCI's) distributed supercomputer (DAS) project systems to provide cohesive support for the past 19 years to computer science researchers in DAS project participant institutions.

SETUP BY A SINGLE ENTITY

All DAS system generations were set up by ASCI—a formal and accredited entity established in 1995 by Dutch universities to stimulate research collaborations. ASCI's steering committee develops a vision for each generation that is based on the current research agenda, writes the grant proposal, and sets up the system in line with this vision. Because of this central organization, each DAS generation can be set up by implementing a single clear vision instead of combining different resources in ad hoc ways. Running the same system software on each system, for example, greatly increases the distributed system's cohesiveness and drastically simplifies DAS maintenance overall, which in each generation requires only 0.5 full time–employee (FTE) hours.

DESIGN FOR INTERACTIVE EXPERIMENTATION

All DAS system generations were designed for computer science research, with an emphasis on interactive distributed experiments. DAS users have access to the entire distributed system and can allocate multiple clusters at the same time. In contrast to production systems, which aim to maximize system use, DAS aims to optimize system availability. Unlike the DAS system, production systems cannot handle the co-allocation of identical resources, which limits their usability.

STIMULATING NOVEL APPLICATIONS

Researchers have consistently used DAS systems to validate new techniques, which has triggered novel research paths in multimedia, model checking, and the Semantic Web. The DAS steering committee encourages collaborative application research on DAS, as long as it does not use DAS for production runs. Each new DAS system generation has stimulated researchers to find novel ways to meet challenges. Notable examples are in video and image retrieval and in astronomy.

VIDEO AND IMAGE RETRIEVAL

A challenge for researchers in video and image retrieval was the lack of programming tools for novices. DAS-3 researchers developed user-transparent programming models that hid the difficulties of parallel implementation from their users, while supporting easy-to-use grid execution.[1] This work led to research on image and video retrieval[2] that won the Most Visionary Research Award at the 22nd Conference on Artificial Intelligence (AAAI 2007) and research on the detection of supernovae candidates in telescopic image data that won the 1st International Data Analysis Challenge (DACH 2008) for Finding Supernovae.

The key challenge in visual retrieval is to look only at the pixels to understand what is happening in the image and where. The standard approach involves processing large amounts of

the Internet of Things, for example, must be turned into massive virtualized data stores. The virtualization and integration of wireless networks into these systems and platforms is the next challenge, which DAS-5 is already aiming to meet by allowing a new layer of network virtualization.

Big data

Big data support requires both new information-processing technologies (for example, machine learning and semantic analysis) to address semantic issues and new high-performance technologies to handle large-volume and dynamic streams of data.

APPLICATION AREAS

We are confident that the newest generation of DAS systems will continue the trend of highly productive collaboration with application scientists. With the performance of its combined clusters coming nowhere near that of top-500 supercomputers, this

imagery to extract multiple color, shape, texture, and motion features and to use supervised learning of labeled examples to convert these to semantic labels, like "sheep dog," "bowling alley," and "teenager." Examples of video and image retrieval innovations are new representations for video and images using color invariants, smart feature pooling, GPU-specific kernel computation, and harvesting training examples from the Web. Recently, object localization in images, deep learning, and video-event recognition has been added to the repertoire.

Algorithmic innovations from DAS-3 and DAS-4 researchers resulted in leading benchmarks for accuracy in video and image retrieval. Researchers won the National Institute of Standards and Technology (NIST) Text Retrieval Conference Video Retrieval Evaluation (TRECVID) competition for concept detection and event recognition six times, as well as contests such as the ImageCLEF photo annotation task in 2009; ImageNet object classification with localization in 2011; Pascal visual object classes (VOC) classification with localization in 2012; and ImageNet object detection in 2013. (ImageCLEF is the image-retrieval laboratory of the Cross Language Evaluation Forum [CLEF].)

ASTRONOMY

DAS systems have been and continue to be useful in evaluating architectures and algorithms for astronomy applications, particularly in analyzing radio-astronomy signals. An extensive study of the impact of auto-tuning for a de-dispersion astronomy program required experimentation with NVIDIA and AMD GPUs and Intel Xeon Phi accelerators, and with datasets from different telescopes. The experiments required a huge number of short runs to auto-tune several OpenCL parameters, such as the optimal number of work items or registers, in a range of scenarios.

The cluster at ASTRON, the Netherlands Institute for Radio Astronomy, was used extensively to prototype applications that exploit new accelerator technologies. These applications, which included two GPU correlators, a beam former, an imager, and a pulsar pipeline, are now used in production by the Low-Frequency Analysis and Recording (LOFAR) telescope, the largest radio telescope in the world.

Other projects, such as the distributed high-performance Astrophysical Multipurpose Software Environment (AMUSE), required astrophysics simulations. In 2010, a GPU cluster was added to the DAS-4 Leiden site funded by a separate grant from the Netherlands Organization for Scientific Research (NWO). Work produced through this additional cluster ultimately led to the experiment to simulate the Milky Way Galaxy on 18,600 GPUs of the US Oakridge National Laboratory's Titan supercomputer—work that was nominated for the Gordon Bell prize in 2014.

References

1. F.J. Seinstra et al., "High-Performance Distributed Image and Video Content Analysis with Parallel-Horus," *IEEE MultiMedia*, vol. 14, no. 4, 2007, pp. 64–75.
2. K.E.A. van de Sande, T. Gevers, and C.G.M. Snoek, "Evaluating Color Descriptors for Object and Scene Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, 2010, pp. 1582–1596.

successful collaboration might seem puzzling. Even though ASCI decided early on to ban production runs during the day, the last three DAS system generations have attracted participation and co-funding from new partners in multimedia for DAS-3, astronomy for DAS-4, and e-science for DAS-5.

Table 3 shows the impact of research in these new domains. From this, we can infer that DAS-5 will be attractive for application developers as they research novel directions. Indeed, numerous application developers first used DAS systems for prototyping and then migrated to production supercomputers. We have found that DAS has several advantages for applications.

Fast and easy experimentation

Experiments with parallel algorithms are not bogged down by complicated

TABLE 3. Research application areas for DAS system users and research impact.

Application area	Focus	Impact
Multimedia	Video and image retrieval	3,500+ citations, winner of the NIST TRECVID benchmark (2008–2010 and 2013–2015), and best paper awards from ACM CIVR (2009 and 2010) and *IEEE Transactions on Multimedia* (2012)
Astronomy	Radio signals, astrophysics	Nominee for Gordon Bell Prize 2014
Climate	Models of sea-level rise	Enlighten Your Research Global Award at SC 2013

ACM CIVR: ACM Conf. Image and Video Retrieval; NIST: National Institute of Standards and Technology; TRECVID: Text Retrieval Conf. Video Retrieval Evaluation

procedures or grant proposals. Researchers conducting work in multimedia, astronomy, bioinformatics, the Semantic Web, and distributed model checking have used DAS systems to easily and quickly develop and evaluate algorithmic alternatives, which were later ported to production platforms. For example, the computational characteristics of the multiphysics simulations, which researchers published in *Nature*,[11] could not have been analyzed without DAS availability.

Distributed testbed

DAS-4 researchers used Ibis to efficiently run the parallel ocean program on multiple clusters by optimizing its load balancing for wide-area distributed systems, similar to earlier work with DAS-1. Climate modelers are now using this code on European production systems, and the research won an Enlighten Your Research Global Award at the 2013 International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2013). This load-balancing optimization enabled further experimentation using several top-10 supercomputers in the US, UK, and Germany—all connected with dedicated optical paths.

Modern hardware

Researchers have the benefit of using modern hardware not yet available on production platforms. For example, climate modelers and astrophysicists used DAS-4 accelerators to study how these new components affected their research. Their results drove others to extend a Dutch supercomputer with accelerators for production runs.

Interest in DAS systems has grown steadily for almost two decades. Whereas large-scale systems often struggle to secure long-term funding, ASCI's strategy has always been to maintain cohesiveness and a modest system size. We believe these choices have influenced the growing interest in DAS systems; since its inception, each DAS system generation has attracted new communities and co-funding partners. Each new generation's design has incorporated WANs through close alignment with SURFnet, which funded these networks.

Studying distributed systems and wide-area networking hand in hand has yielded several innovative results from a hardware budget of €1.5 million for each system. Centralized control has been a key factor in this maximum return on investment; without it, the given budget would be lost to fragmented systems management for maintaining numerous local facilities and ad hoc resource sharing. The homogeneous DAS system design drastically simplifies maintenance.

There have been design tradeoffs as well. For example, unlike Grid'5000, DAS-5 does not provide user access to the computational nodes' bare hardware, which enables running all system software stacks. ASCI felt that this approach would complicate design and system management and instead opted to provide nodes with preconfigured stacks, which can be modified if needed. This precluded a few projects, notably new OS designs and low-level intrusion-detection mechanisms. Also, DAS does not support large-scale experiments; to offset that limitation, researchers use other systems (typically international) as needed.

Central organization and system cohesiveness will continue to be DAS system cornerstones because of their importance in supporting award-winning research and obtaining long-term funding from highly competitive programs. Exploring evolving large-scale computing infrastructures like datacenters and clouds will require research infrastructures with successive new generations that can meet pressing research problems. We are confident that DAS-5 and its future designs will meet those needs.

ACKNOWLEDGMENTS

This article is based on a keynote lecture given by Henri Bal at Euro-Par 2014 upon receipt of the Euro-Par Achievement Award. Hundreds of people have contributed to the DAS project by using its infrastructure to conduct novel research. In particular, we thank Andy Tanenbaum, Bob Hertzberger, and Henk Sips, ASCI board members who took the initiative for the first DAS system. We also thank Kees Verstoep, who coordinated the systems management of all DAS generations, and we are grateful to NWO, SURFnet, VL-e, MultimediaN, the Netherlands eScience Center, and the COMMIT/ public-private research community for their funding. Finally, we commemorate the late Lex Wolters, who was a steering committee member for several DAS system generations.

REFERENCES

1. R. Bolze et al., "Grid'5000: A Large-Scale and Highly Reconfigurable

Experimental Grid Testbed," *Int'l J. High Performance Computing Applications*, vol. 20, no. 3, 2006, pp. 481–494.
2. R.A.F. Bhoedjang, T. Ruhl, and H.E. Bal, "User-Level Network Interface Protocols," *Computer*, vol. 31, no. 11, 1998, pp. 53–60.
3. A. Plaat, H.E. Bal, and R.F.H. Hofman, "Sensitivity of Parallel Applications to Large Differences in Bandwidth and Latency in Two-Layer Interconnects," *Proc. 5th Int'l Symp. High-Performance Computer Architecture* (HPCA-5), 1999, pp. 244–253.
4. T. Kielmann et al., "MagPIe: MPI's Collective Communication Operations for Clustered Wide Area Systems," *Proc. ACM SIGPLAN Symp. Principles and Practice of Parallel Programming* (PPoPP 99), 1999, pp. 131–140.
5. C.A. Pickover, *The Math Book: From Pythagoras to the 57th Dimension, 250 Milestones in the History of Mathematics*, Sterling, 2009.
6. R.V. van Nieuwpoort et al., "Ibis: A Flexible and Efficient Java based Grid Programming Environment," *Concurrency and Computation: Practice and Experience*, vol. 17, nos. 7–8, 2005, pp. 1079–1107.
7. M. Jelasity et al., "Gossip-Based Peer Sampling," *ACM Trans. Computer Systems*, vol. 25, no. 3, 2007; doi: 10.1145/1275517.1275520.
8. J.A Pouwelse et al., "TRIBLER: A Social-Based Peer-to-Peer System," *Concurrency and Computation: Practice and Experience*, vol. 20, no. 2, 2008, pp. 127–138.
9. T. DeFanti et al., "TransLight: A Global-Scale LambdaGrid for e-Science," *Comm. ACM*, vol. 46, no. 11, 2003, pp. 34–41.
10. J. Urbani et al., "Scalable Distributed Reasoning Using MapReduce," *Proc. 8th Int'l Semantic Web Conf.* (ISWC 09), 2009, pp. 634–649.
11. S.F. Portegies Zwart and E.P.J. van den Heuvel, "A Runaway Collision in a Young Star Cluster as the Origin of the Brightest Supernova," *Nature*, vol. 450, 2007, pp. 388–389.

ABOUT THE AUTHORS

HENRI BAL is a professor of high-performance distributed computing at VU University Amsterdam (VU). His research interests include programming environments for parallel, distributed, and smartphone systems. Bal received a PhD in computer science from VU. He is a member of IEEE, ACM, and the Academia Europeana. Contact him at bal@cs.vu.nl.

DICK EPEMA is a professor of distributed systems at Delft University of Technology. His research interests include scheduling in large-scale distributed systems and online social networks. Epema received a PhD in mathematics from Leiden University. He is a member of IEEE and ACM. Contact him at d.h.j.epema@tudelft.nl.

CEES DE LAAT is a professor of system and network engineering at the University of Amsterdam. His research interests include optical and software-defined networking, workflows, big data, e-infrastructures, and systems security and privacy. De Laat received a PhD in physics from the Delft University of Technology. He is cofounder of the Global Lambda Integrated Facility and a member of IEEE and ACM. Contact him at delaat@uva.nl.

ROB VAN NIEUWPOORT is director of technology at the Netherlands eScience Center. His research interests include efficient computing and scalable parallel programming models, and their application in various research disciplines. Nieuwpoort received a PhD in computer science at VU. Contact him at r.vannieuwpoort@esciencecenter.nl.

JOHN ROMEIN is a senior researcher at ASTRON, the Netherlands Institute for Radio Astronomy. His research interests include high-performance computing for radio-astronomical applications, with a particular focus on accelerators. Romein received a PhD in computer science from VU. Contact him at romein@astron.nl.

FRANK SEINSTRA is director of the eScience program at the Netherlands e-Science Center. His research interests include advanced distributed cyberinfrastructures (jungle computing) and their application in various research domains. Seinstra received a PhD in computer science from the University of Amsterdam. Contact him at f.seinstra@esciencecenter.nl.

CEES SNOEK is an associate professor of computer science at the University of Amsterdam. His research interests include video retrieval and image retrieval. Snoek received a PhD in computer science from the University of Amsterdam. He is a Senior Member of ACM and IEEE. Contact him at cgmsnoek@uva.nl.

HARRY WIJSHOFF is a professor of computer science at Leiden University. His research interests include restructuring compilers, irregular computations, and parallel algorithms. Wijshoff received a PhD in computer science from Utrecht University. Contact him at h.a.g.wijshoff@liacs.leidenuniv.nl.