Autonomous IT: Outsourcing Operations to Machines

BY DAN FARBER

This issue of Release 1.0 reflects our increasingly close collaboration with CNET Networks: We lured Dan Farber, longtime editor in chief of ZDNet, to write it. He has been with CNET since 2000, when it acquired ZDNet. But we would have picked Farber anyway; this topic is one worthy of his talents as a technology journalist for more than 20 years. When we discussed this issue, he said with a gleam in his eye: “We tend to equate technology with automation, but we are still in the Stone Age. The next era of computing is getting underway, and it’s all about machine-to-machine communication, rather than IT administrators turning dials and knobs to keep the systems running,” he says.

Farber has long been intrigued by the meta-trends in this industry. And of course, not only does he understand this stuff; he can explain it. Even though this market hardly exists as yet, its outlines are clear:

As systems get more complex, we have to start trusting them to manage themselves; it’s simply beyond the capability of human beings. We can’t clone humans fast enough, but we can clone smart, aware machines if we can make the systems themselves manageable – by making their states and capabilities explicit. Yet there’s more to this than theory and technology; the marketplace comprises living, breathing companies with histories and tactics. The companies that control the platforms – notably IBM and Microsoft – are attempting to put more intelligence into the infrastructure, while those with less market power are mostly working on smart middleware to control the platform components from outside. But in the long run, all these approaches will benefit from increasing standardization, which will make everything easier to manage – and reduce any one vendor’s market power.

– Esther Dyson
You are the CIO of a large manufacturing corporation. Your budget for IT is $35 million this year. Your company has more than 12,000 employees, spread out across 30 facilities in 25 countries. Currently, 450 people work in IT, managing 550 applications (mostly home grown), with 10 operating systems versions, thousands of servers, 15,000 PCs and 3,500 mobile devices.

Your corporate mandate, driven from the CEO and the board of directors (who are petrified by Sarbanes-Oxley and have prioritized spending on compliance), is to lower the cost of IT by 25 percent in the next 18 months, and at the same time deploy upgrades to CRM and reduce out-of-stock parts in the supply chain from 5 percent to 2 percent by year-end. The company also wants to fire up new distribution and development locations in Asia.

You call upon executive staff, business division heads, consultants and peers, but you are hard pressed to come up with more than 10 percent in cost savings, and most of that is gained through server consolidation and outsourcing to lower-cost offshore shops. You’ve started a few Web services deployments and are evaluating provisioning and configuration management tools, but you are cautious about betting heavily on newer technologies. You can’t reduce staff, the biggest IT cost center, without compromising IT service levels.

IT Nightmare: Cost and Complexity

This is the situation in which CIOs, IT executives, system administrators and even small business owners find themselves today. Technology is vital and necessary to the conduct of business, but it’s costly and increasingly complex. Businesses are demanding more control, accountability and leverage of existing infrastructure.

As described by Francis Heylighen, a research professor at the Free University of Brussels who focuses on the evolution of complexity, information systems today “depend on so many modules, sources of data, network connections and input and output devices that it has become impossible to predict or control their interactions. The
result is software full of bugs, corrupted data, security holes, viruses, and other potentially catastrophic side effects. Moreover, systems have become so complex that the human mind simply no longer can learn or remember all procedures needed to use them. Add to this constant change in hardware, software, protocols, data and user expectations, mix it well, and you have a recipe for chaos.”

Dealing with that complexity and chaos is the dominant battlefront today for both vendors and their customers. Despite the steady decline in the cost of hardware and software over the last 25 years, the operational cost of running an IT environment has gone in the opposite direction. According to research firm Gartner, 60 to 70 percent of IT expenditures are dedicated to staff assigned to operations and maintenance tasks. IDC estimates that companies worldwide spent nearly $174 billion in 2003 for ongoing operations and management of their existing IT environments. Merely applying a software patch to a PC can cost $150 per patch, according to the Yankee Group.

Recently, American Airlines and US Airways were forced to ground hundreds of flights after an employee of Electronic Data Systems, which manages and hosts the applications for the airlines, used a “bad command” about a cargo shipment, resulting in a glitch in the database used for baggage handling, scheduling crews and dispatching flights. Just identifying the problem required a small army of people at EDS.

Most of the IT systems installed in corporations today were designed with the implicit notion that a legion of human beings would be on hand to tell the systems what to do. Continuing to add more IT people to mask the complexity and identify the fault lines of IT infrastructure is economically untenable for businesses. The increasing complexity of the IT environment also creates immense management and scaling challenges: Adding a single component that interacts with 100 other components within a data center, for example, requires 100 more discrete interactions. You want to increase the potential complexity of your business without increasing the complexity of managing the underlying systems, following the rule of Occam’s razor: Be no more complex than you have to be.

Sun CEO Scott McNealy says: “We’re at roughly the same state now as the telephone industry when it still had switchboards and operators who manually connected calls. If we don’t develop a system that scales better than the equivalent of automatic switching we’ll have to retrain everyone under the age of 30 to work as system administrators.”
Curt Cotner, an IBM Fellow and a relational database pioneer, agrees: “If you look out over the next 10 years, you’ll want to know who is doing the best job of reducing people costs. Outsourcing won’t eliminate cost in the magnitudes required. The day of different vendors competing on hardware performance and price will end. The next battle ground is on people costs: How much can you automate and how much intelligence can you get into the software and hardware?” In other words, let the machines do the work.

**The slow road to IT automation**

As McNealy suggests, we can’t scale computing infrastructure – or address the confluence of cost and complexity – without addressing the issue of IT automation. Over the next 10 years, solving the scaling and automation problems will be a focus of companies large and small.

The ultimate in IT automation is the “lights-out” data center, in which IT infrastructure is **autonomic**, managing itself by embedded digital sense-and-response mechanisms, similar to the way the human nervous system functions. (See page 15.) A single IT administrator could be responsible for hundreds of thousands of servers.

While humans are well suited to creating blueprints for how systems should function, machines are far better suited to manage those complex operations in real time. They don’t fall asleep, suffer from circadian rhythms, get bored or ask for pay raises; they are **not** creative. Any time a server workload reaches a level that threatens a service-level guarantee, for example, the self-managing IT environment could follow policies created by a human and automatically add a server, or pull the resources from servers working on lower-priority tasks and configure them with the proper software and connectivity. Embedded policies and heuristics could automatically address problems such as component failures, security breaches and workload spikes.

Some of those self-managing capabilities exist today, as we’ll discuss below, but the “lights-out” data center that takes all IT personnel out of the equation is far more dream than reality. “If you correlated all information about the IT infrastructure, and knew all potential outcomes, you could get to a zero-person data center. But the reality is that we don’t know all the problems and permutation of problems that could happen [especially in a world of proliferating standards]. We are somewhere between everything being manual and being people-less,” says Robert LeBlanc, general manager of IBM’s WebSphere platform.
According to LeBlanc, 90 percent of all actions that can occur in a data center are routine in nature. It’s the 10 percent that doesn’t fall within normal ranges that is the Achilles’ heel of IT automation and that will continue to require human intervention.

Don LeClair, senior VP in the office of the CTO at Computer Associates, defines the lights-out data center as reducing costs by 90 to 99 percent, including costs for helpdesk support per employee, maintenance and upgrade costs per desktop, administrative costs per server, administrative costs per terabyte of storage, and the costs (in time) to provision a server. Those radical cost savings will be achieved incrementally over the next decade, as enterprises deploy the new and refurbished technologies and concepts described below.

Climbing onto Dry Land

As with any other major overhaul, IT automation is an evolutionary, not a revolutionary, process. The evolution forces both vendors and their customers to create and adopt new models for building IT infrastructure. The Internet took nearly 30 years to reach mass adoption, and depended on the standardization and interoperability fostered by protocols such as TCP/IP and HTTP. Internet protocols allowed users to communicate without borders, abstracting away the difference between servers and other infrastructure components.

IT automation requires the same abstraction, allowing a wide variety of specific resources to interoperate fungibly. Once the resources are fungible and can be described and manipulated in standard ways, machines can directly manage and enforce policies to run IT infrastructure efficiently.

IT automation is just getting its legs, climbing onto dry land, but the pieces to bring it into the mainstream within the next decade are converging. Just as standards such as TCP/IP catalyzed the development of the Internet, Web services and service-oriented architecture (SOA), virtualization, grid computing, policy-based management and open standards provide the framework for building a world of automated IT. Combined, they commoditize IT infrastructure, lowering the cost and complexity of delivering and supporting applications for businesses and homes.

**Web services and SOA:** The evolution of open standards – such as XML (eXtensible Markup Language), SOAP (Simple Object Access Protocol), and UDDI (Universal
Description Discover and Integration) – for dynamically interacting Web applications has been broadly supported by the two major development platforms, Microsoft’s .NET and Sun’s Java.

Web services are interface standards that allow applications to communicate with each other, independent of the underlying platform and programming language. Technically, Web services are reusable software interfaces that describe operations that can be accessed over a network through XML messaging. (See Release 1.0, September 2001 and October 2001.)

Related to Web services, service-oriented architecture (SOA) is a component model framework for the development and deployment of Web services, and has emerged as a key strategy for building applications. (See Release 1.0, December 2002.) “The mandate for CIOs is to move to SOA at NASCAR speed to get the benefits of reduced complexity and cost,” says Prasad L. Rampalli, VP of finance and enterprise services and director and chief architect of architecture and integration platforms at Intel.

Rather than each vendor coming up with its own way of automating infrastructure, SOA and Web services allow vendors to use a common way to expose software functionality. As a result, changes to services that compose an application or business process can be made easily, reducing labor costs and creating more flexibility in delivering IT services. In addition, Web services help bridge the gap between business people and IT: A business person can describe a business process in plain English and an IT person can associate the description with appropriate Web services using simple, visual design tools.

For example, if you are developing a printing application as part of a business process to print statements at the end of the day, a service-oriented infrastructure application can expose both the operational functions (such as printing) and management parameters (such as availability, service-level status, and performance). With this simplified interface and exposed information, a Web service application that collects status information from all the printers can execute pre-defined policies to determine the optimal place to print in an environment with multiple applications requesting print services. A human does not need to get involved, and all the applications (including your statement-processor and other departments’ applications) get printing functions allocated optimally according to their needs and priorities, and the printers’ capabilities, across multiple environments and in changing conditions. In this case, the priorities given to the various applications (or their
users) are set up as policies by IT departments, who reflect management’s wishes. The policies can, of course, vary by time of day or month or any other parameters.

In the more “radical” version of IT automation, the priority information might be embedded in the applications – as a budget, for example – and the printers would bid against one another for the jobs, following their own internal states of busyness and other metrics. This vision of market-oriented, self-organizing resources works best either in a closed system, where the “owner” defines the currency, the competing requirements and the resources to use, or in a totally open system, where the currency is actual money.

Right now, the middleware manager scenario is the default; it’s the approach favored by vendors who can develop tools to manage third parties’ applications – either by determining their status or by reading their exposed information. It’s the appropriate model for a start-up or a second-tier player that doesn’t control the applications or their design. The “radical” approach, espoused by Microsoft and IBM, puts more of the intelligence into the applications – either the requesting business processes or the service-providing infrastructure apps – and lets them organize themselves. (See Pages 15 and 30.)

Every CIO is developing a Web services/SOA strategy, but like those for grid computing, Web-services standards are still developing, and issues around business-process schemas, reliability, performance and security still need to be resolved (see Release 1.0, December 2002). Nonetheless, the train has left the station. Gartner predicts more than 60 percent of enterprises will use SOA as the “guiding principle” of their IT infrastructure development by 2008.

**Virtualization:** The concept of virtualization, such as partitioning of disk space to create multiple execution environments (virtual machines), has been around for decades on big iron systems. Virtual machines are used to consolidate server workloads, allowing a single server to run multiple operating system and application instances. Virtual infrastructure works by inserting a layer of abstraction between hardware resources and the software that runs on them. The benefits of virtualization include increased hardware utilization rates, easier manageability and faster, smoother response to changes in business processes and applications. With virtualization, systems can be reconfigured and services provisioned under the covers; developers/operators don’t have to manually write shell scripts, Perl or Java code to make changes.
Today, virtualization is moving downstream into low-cost hardware platforms. For example, Palo Alto-based VMware’s software virtualizes Intel-based systems into mainframe-class virtual machines, allowing for server reconfiguration on the fly. Major storage vendors also offer virtualization, which aggregates storage assets into pools that can be centrally managed and delivered to applications.

However, fully automating IT will require more “extreme” virtualization, in which all IT infrastructure resources – processor cycles, storage bits, and network throughput – are virtualized into a single pool that can be dynamically allocated to applications.

**Policy-based management:** IT infrastructure is often an unwieldy, unreliable tangle of software and hardware, each component with its own set of quirky behaviors and configuration requirements. Integrating a mixture of legacy systems with future IT investments can be an exercise in futility. Management and integration software – usually proprietary and consigned to islands of IT infrastructure – from dozens of vendors monitors the behavior of desktops, servers, storage systems, networks and applications, and correlates the collected data against databases of accepted behavior to identify problems. That’s not sufficient to automate IT.

IT infrastructure must be integrated and “instrumented” so it can be controlled by policies expressing the capabilities, requirements, and general characteristics of IT components that allow the individual components themselves to adapt to changes in the environment. Policies define the operational procedures and sets of actions that carry out some function or process independently of the components used and the specific conditions under which a task is carried out. For example, when a server reaches 80-percent utilization the management software could initiate a process to provision another server to balance the workload or give specific applications or users preferential access to bandwidth.

Policy-based automation is the critical middleware between the applications and business processes and the pool of resources distributed across an enterprise or beyond. It’s like a conductor, orchestrating the discrete components embedded in different layers into a harmonious workflow. When physical resources – regardless of operating system, hardware or application – are abstracted via virtualization, business processes and Web services can be mapped onto the pool of resources and delivered autonomously.

From a user perspective, it’s about specifying what to deliver rather than how to deliver an IT service. However, every company that offers policy-based management
tools today has its own XML-based language for operational controls in SOA environments. In the next few years, we can expect a more standardized policy-management semantic platform, which will improve interoperability among different vendor solutions and lower the cost to deliver IT automation.

**Grid computing:** Extreme virtualization is tightly bound to grid computing, which first surfaced in the mid-1990’s as a way to describe a distributed computing architecture for high-performance computing applications. (See **Release 1.0, November 2002.**) Grids use virtualization to make distributed, rather than just local, compute resources appear as a single pool of resources. Ian Foster, grid computing pioneer, senior scientist and head of the Distributed Systems Lab at Argonne National Laboratory, and professor of computer science at the University of Chicago, describes grid computing as a “system that coordinates resources that are not subject to centralized control, using standard, open, general-purpose protocols and interfaces to deliver non-trivial qualities of service.”

Within a grid, a wide variety of operating systems and hardware is turned into a single, virtual computer. Grids will provide more pervasive IT automation by harnessing distributed hardware and software resources, virtualizing them, and managing them autonomically in a decentralized, heterogeneous service-oriented framework. Instead of requisitioning a dual-processor server system and so many gigabytes of Fibre Channel-based storage to run an Oracle database application, an administrator thinks about hardware needs in terms of service-level objectives, such as sub-second response time for the database or 99.999-percent uptime.

To date, grid computing has been used to coordinate computing resources from multiple owners to handle a single large scientific task, such as the SETI@Home project, which harnesses 5 million PCs to search for deep-space radio signals from extraterrestrials, or IBM’s Butterfly.net, which uses a grid for a multiplayer game network. Grid.org, a website for large-scale research projects powered by Austin-based United Devices’ grid computing solution, harnesses 2.5 million systems in more than 225 countries to deliver in excess of 150 teraflops of power to applications. Using grid.org, the Anthrax Research Project screened 3.57 billion molecules for suitability as a treatment for advanced-stage anthrax in 24 days. The screening would have taken years using conventional methods.

The next phase for grid computing is to apply grid computing and IT automation concepts broadly as a framework for administering commercial enterprise IT infrastructure. This transition to enterprise grids will take place in stages. It has already
started with grids inside the firewall, which allow sharing of resources within a corporation or department. Intel, for example, has built a grid for microprocessor development that harnesses compute cycles from 35,000 workstations and servers, divided into 275 pools. The Intel grid runs 5 to 6 million jobs, or processes, per week. Each month about 1,500 platform problems, such as fixing an improper configuration, are identified and resolved automatically via embedded policies, according to Ravi Subramaniam, manager of distributed computing solutions at the company.

The $30-billion company saves $75 million annually using its internal grid, but its software is not off-the-shelf and the system is both “labor-intensive to set up and run and brittle,” Subramaniam says. In addition, grid implementations today provide only limited guarantees of availability and security. However, the development of standards and tools for building enterprise grids will overcome those shortcomings in the next few years (seebox, page 20).

Other commercial grid applications today include reservoir modeling for petroleum exploration, actuarial analysis in the insurance industry, structural analysis in aerospace and automotive companies and in-silico testing by pharmaceutical companies. Several private companies – including Avaki, Data Synapse, Platform Computing and United Devices – specialize in grid software and have businesses serving Fortune 1000 companies.

In addition, more mainstream applications are becoming grid-friendly: Platform Computing, headquartered in Toronto with $50 million in revenue for 2003, has an adaptor for Microsoft Excel that grid-enables compute-intensive worksheets. Oracle 10g, the company’s flagship database, allows database administrators to virtualize, pool, and provision data to a grid, improving performance and increasing hardware utilization. In addition to cost savings through higher hardware utilization rates, Oracle 10g also has automated functions for building out and managing 10g grids.

So called “extragrids” will enable enterprises to share computing resources with trusted partners and suppliers in a kind of federated grid. The final stage, which will take a decade to evolve, is the utility grid, in which compute resources are delivered on demand in a pay-for-use model with distributed, rather than centralized, control.

Utility computing grids will require maturing of various technologies and standards as well as business models for allocating and charging for resources and assigning liability. They will also require cultural adaptation: Most CIOs are not ready to give up control of their IT infrastructure to utility service providers, nor are they ready to
entrust their data and core intellectual property outside their firewalls. (see release 1.0, december 2002.)

Tapping into the external grid for services will require service-level guarantees that contain notions of trusted, bonded resource providers. Just as federated identity networks will be governed by common operating rules and legal covenants such as standardized conflict resolution procedures and mutual confidence parameters, grid-enabled IT infrastructure will need to develop similar enforceable guidelines.

Nor are those CIOs ready to fully trust machines to operate IT infrastructure, whether from an “intragrid” that uses distributed assets inside the corporation, an extragrid, or from a self-organizing system. “Trusting machines completely to provide ‘five nines’ of reliability or to keep an enterprise safe from a denial-of-service attack isn’t going to happen any time soon,” says David Nelson-Gal, senior VP of engineering at Interwoven and former VP at Sun for the N1 Grid platform. “When you first swiped a credit card at a gas pump, you wanted a receipt. [That’s less true] now since more trust is built up. You’re not worried about having tons of cash siphoned off. In similar ways, customers want to see what automation is doing and to have control points. It will evolve to where it’s more hands-free over time.”

Standards, especially open ones: Although Web services standards are percolating nicely, many other standards are needed for IT automation to succeed. Storage vendors, for instance, haven’t yet devised a consistent definition for storage assets so that they can be easily virtualized. Standards for cross-vendor management are also lacking: “The culture of management vendors is to compete on depth of product integration, flying in the face of cross-vendor integration,” says Corey Ferengul, senior VP at META Group. The protectionist attitude of vendors encourages islands of vendor-specific automation, but several standards bodies — such as the Distributed Management Task Force (DMTF), Global Grid Forum (GGF), and the Organization for the Advancement of Structured Information Standards (OASIS) — are working together to establish new management standards.

Until customers lobby more strongly for IT automation, vendors will continue to give lip service to cross-vendor management standards in the short term. But in the longer term, dodging more universal standards is a dangerous strategy for automation vendors. Over the next few years, vendors without a business to protect will do what Linux has done in the operating system world: create a new model based on open standards, if not open source. We’ll look at a few of those potentially disruptive companies that are beginning to emerge.
The Next Big Thing

For enterprises, the benefit of these converging technologies/approaches is reducing labor cost, increasing reliability and squeezing more out of existing physical resources. In addition, companies can shift to more variable and predictable cost scenarios, paying only for what they use. Most importantly, IT infrastructure is more adaptable and responsive to changes in business priorities without fielding an army of IT administrators.

The opportunity just for automating server management is huge. IDC Research estimates that $95 billion will be spent this year – mostly on labor costs – managing the 20 million servers humming away in data centers worldwide. According to Diane Greene, CEO of VMware, only about 1 percent of these 20 million servers have been virtualized. As a result, server utilization rates average around 10 to 15 percent today.

According to Gartner, by 2008 companies that don’t virtualize their servers will spend 25 percent more annually for hardware, software, labor, and rack space for Intel-based servers, and 15 percent more for RISC servers. IDC predicts that automated server management, provisioning and virtualization will be a $5-billion market by 2008.

Mendel Rosenblum, chief scientist at VMware and an associate professor of computer science at Stanford, predicts that virtualization software will ship on every server in the next few years. “Provisioning, patching or consolidating servers, whether a box, blade, physical or virtual server, will require only minimal human intervention. The administrator per server ratio, measured in OS images or processors, will continue to decrease, lowering the cost of supporting a data center,” he says.

IDC predicts that the nascent market for grid computing will exceed $12 billion by 2007 across commercial and high-performance technical computing markets.

Islands of automation

As Forrester Research analyst Richard Fichera notes, overarching enterprise-wide IT automation solutions are impractical today. “The early stages of the adoption cycle have shown us that many organizations are struggling with this strategy at a number of levels other than the underlying technology, including the organizational issues, the imposing tasks of overlaying an adaptive infrastructure on top of a legacy environment, and the setting of internal priorities and investment levels. In some cases, promising product technologies have come to a sudden halt against these barriers.
We have to rethink the essential strategy for customer deployments and go back to what amount to islands of utility functionality – automated infrastructure and application domains dedicated to a specific application or group of applications."

In addition, each vendor’s solution is unique, based on hodgepods of existing, emerging and “faux” standards, potentially locking in customers. With different policy-management or grid approaches, for example, integrating data centers or other assets across an enterprise becomes a complicated mapping exercise.

Most solutions to date have focused on executing a few tasks, such as installing software patches, tracking assets, or monitoring system activities. With the emergence of Web services, the market opportunity has moved to multifunction solutions that orchestrate a set of related tasks across and enterprise, with an understanding of the dependencies and logic. These are still islands of automation, but they can move beyond a single data center or outside the corporate firewall.

**A rose by any other name**


Despite the R&D dollars being spent at the larger firms, much of the innovation in orchestrated IT automation has come from smaller companies, and the billion-dollar vendors have been scooping them up over the last few years. In May 2003, IBM acquired Think Dynamics, a 36-person company with software that lets customers create policies that can then automatically reallocate computing resources on the fly to meet spikes in demand or to react to failures in heterogeneous environments.

Now Think Dynamics’ software is packaged as IBM’s Tivoli Intelligent Orchestrator.

In August 2003, Sun acquired CenterRun, which automates provisioning applications, including configuring, distributing and setting up packaged and custom updates, patches and applications in mixed environments. CenterRun, now integrated into Sun’s N1 Grid Provisioning System, also enables simulation of application changes before deployment.
In February of this year, HP acquired Consera Software, which develops software that allows users to customize and create workflows to automate IT management tasks. The product, now named HP OpenView Service Delivery Designer, enables companies to build graphical models that represent business services mapped to underlying IT components. In April, HP also acquired Novadigm, which developed policy-based management software for automating IT infrastructure. BMC Software acquired Marimba, which has server and desktop provisioning and change management software, in July.

In January, EMC acquired Palo Alto-based VMware for $635 million in cash. VMware’s software allows multiple operating systems to run on the same X-86-based physical hardware and supports allocation of system resources to any operating system as needed. The virtual server infrastructure software reduces hardware costs by increasing virtualization, and lowers operational costs by enforcing configuration and performance policies, consolidating server-based applications and automating provisioning of servers.

The company is recognized as the de facto industry standard for X-86-based server virtualization, with over 5000 customers and strategic partnerships with Dell, HP, IBM, and Unisys. In October Microsoft will be shipping its Virtual Server 2005, also for X-86 systems, but it provides a more basic feature set than VMware.

VMware and Microsoft won’t necessarily carve up the X-86-based system virtualization between themselves. Acton, MA-based, stealth-mode startup Katana Technology was funded with $20 million last year from Matrix Partners and Highland Capital Partners to develop a product that virtualizes discrete server components such as processors and memory and I/O on Intel-based hardware. The pedigree of the founders is impressive: CEO Scott Davis was formerly CTO at Mangosoft and technical director for Digital’s industry-acclaimed VAXCluster and VMS Volume Shadowing products, as well as Digital’s Windows NT clustering technology (later sold to Microsoft as the genesis of its “Wolfpack” Cluster Server). Katana chief scientist Alex Vasilevsky was a senior engineer at Avid Technology and Thinking Machines, where he worked its famous parallel supercomputer, the Connection Machine.

While all these brands represent the articulation by the major vendors of the broad outlines of IT automation, at this point they are movie trailers meant to sell customers on coming attractions. We are in a “walk before you run” phase, in which customers are cautious about giving up control to machines and the vendors are just beginning to flesh out more complete IT automation solutions. They all share a
common goal: instrumenting IT infrastructure management so that machines, not humans, do the heavy lifting. The movies are still in production in R&D shops, but the early versions are getting sneak previews by handfuls of customers.

In the following pages, we’ll look at how some of the major vendors are setting up shop to automate all of IT administration, and at some smaller players, including Opsware, UXComm, Sychron, and Enigmatec, trying to get a piece of the action.

In addition, we probe stealth startups Cassatt, which claims to have discovered the Holy Grid of IT automation, and an as yet unnamed company headed by Silicon Valley veteran Rich Miller. Finally, we’ll give an indication of where IT automation is headed longer-term.

Autonomic computing: IBM’s lights-out vision
IBM has gone the furthest in terms of articulating a vision for the ultimate in IT automation with its concept of autonomic computing. The term “autonomic computing,” which grew out of cybernetics, control theory and concepts from self-managing biological systems, was brought to the commercial forefront by IBM senior VP and head of research Paul Horn. In 2001 Horn published an essay, “Autonomic Computing: IBM’s Perspective on the State of Information Technology,” which discussed the need for IT systems to manage themselves and for new components to integrate as “effortlessly as a new cell establishes itself in the human body.” With luck, the “cells” for automating IT are cancer-free!

“The information technology boom can only explode for so long before it collapses on itself in a jumble of wires, buttons and knobs. IBM knows that increasing processor might, storage capacity and network connectivity must report to some kind of systemic authority if we expect to take advantage of its potential. The human body’s self-regulating nervous system presents an excellent model for creating the next generation of computing: autonomic computing,” Horn wrote.

Alan Ganek, VP of autonomic computing at IBM, calls it acting without conscious thinking. “A basketball player charging down the court isn’t thinking about how to constrict the pupil of his eye, or how to elevate his breathing and heart rate,” he says. Rather than hard-coded systems lacking in adaptability or awareness of their environment, autonomic computing is self-configuring (adapting dynamically to changes in the environment), self-healing (discovering, diagnosing and triggering policy/context-based remedial actions), self-optimizing (constantly tuning the envi-
ronment for optimal performance based on business requirements) and self-protecting (proactively anticipating, detecting and protecting against/deterring attacks).

The other vendors have different terms to describe their visions of autonomic computing, which is confusing for customers, and smaller R&D budgets than IBM, but the benefits are the same – masking complexity, increased utilization rates, reduced human intervention, higher reliability and faster implementation of changes.

In an autonomic computing environment, the IT components communicate and collaborate with each other and are managed through higher-level tools that mask their complexity, Ganek says. IBM has developed an Autonomic Management Engine (AME), which implements control loops that process and analyze event information and use policies to govern the managed resources.

“We don’t envision a big building in central Manhattan with computers and no humans,” says Steve White, senior manager, Massively Distributed Systems Group at IBM. “We want to figure out how to apply computers to do the drudgery and for the humans to conceptualize and have level control.”

IBM is trying to make autonomic computing less research project and more reality, infusing its WebSphere software product line and enterprise applications with IT automation features. The company has long served enterprises and itself manages/supports a broad range of silos. Now it is trying to lead by working with partners to integrate its applications with its AME and the Tivoli and WebSphere platforms, and by developing standards to help solve the interoperability problems. Not a bad solution, as long as the standards are its own.

One such proposed standard, developed in conjunction with Cisco, is Common Base Event (CBE), a data format for assisting in problem determination. IT infrastructure components typically have their own unique formats for capturing data about events in log files. If an e-commerce application has a problem, for example, a bunch of IT administrators would have to pore over dozens of log files, each with a different format. As a lingua franca for problem determination, the CBE makes it far easier to diagnose the root cause problems in complex IT environments. IBM has submitted the CBE to the OASIS Web Services Distributed Management Technical Committee.

IBM also led the development of Solution Installation for Autonomic Computing, an XML-based specification aimed at standardizing the installation of business applications. The specification allows developers to capture the dependencies and
prerequisites of a software application, similar to the Data Center Markup Language (see box, page 20) and Microsoft’s System Definition Model (see page 30). For example, an e-commerce application could require a specific server configuration. Solution Installation tools could automatically detect problems and try to fix them.

IBM’s autonomic computing initiative, as part of the company’s overall On Demand computing effort, is dependent on slow-moving standards and slow-moving customers, but it is also affecting IBM’s applications. Cotner points to improvements in the company’s DB2 database as an example of movement toward autonomic computing. A module of the 2006 release of DB2 for z/OS, a mainframe-level system, will automatically set the size of the buffer pool. The module will coordinate the database, operating system and hardware down to the silicon to adjust the buffer size dynamically to prevent overflows that slowdown performance. “Tuning buffer pools is a dark art,” Cotner says. “Very few DBAs [database administrators] can do it well.” So you get a cheaper solution with more consistency.

“Once in place, this code would be smart enough that if some malicious or uninformed user submitted a query with massive [buffer space] requirements and was not recognized as a priority according to the service-level policies, the database would ignore the request. That level of knowledge and auto-enablement is a big leap forward,” Cotner says. However, instrumenting components with triggers and policies that automate actions can suffer from the law of unintended consequences. “DB2 on z/OS has over 6 million lines of code,” Cotner says. “As you collect up all the knobs, dials and levers on products over 25 years, for instance, it gets beyond human comprehension to operate.” As a result, automating the legacy applications – adding new system parameters – also is subject to laws of diminishing returns.

HP and Sun have developed autonomic computing capabilities in their products as well. Sun’s Solaris 10 Unix operating system, for example, has enabled self-healing for CPU, memory, and I/O bus components. Error messages typically intended for humans are instead passed along as telemetry events to software components, which automatically diagnose the underlying fault and initiate self-healing activities, such as deactivating faulty components or sending an alert to an administrator.

HP’s Utility Data Center, which is a hub for centrally managing heterogeneous data center resources and for hosting enterprise applications, has facilities for automatically reconfiguring server environments, such as changing an Apache application server to a Microsoft Exchange server, says Nick van der Zweep, director of virtualization and utility computing at HP.
IT Automation Upstarts

Although the market for IT automation may appear to be coalescing through consolidation, dozens of new companies are still sprouting up. The market is still embryonic, and it is poised to move into high gear in the next 12 months. Newly established players such as Opsware and a crop of venture-backed startups are creating autonomic-like software for managing IT infrastructure. All of them are focused on policy-based automation, orchestrating multi-step business processes and cutting across all vendor implementations and layers in the stack. This new breed of operations-management middleware vendors share similar characteristics. They work with existing infrastructure management and monitoring solutions, such as HP OpenView, CA Unicenter or IBM Tivoli Management Framework. All of them can automate common IT tasks and some can construct high-level, goal-oriented policies that automate IT functions.

The likely scenario for companies profiled here is that they will be acquired by CA, EMC, IBM, HP, Sun, Veritas, and other big consolidators. Part of IT automation is lowering the operational risks, but building IT automation that involves dozens of small vendors is a recipe for disaster. Enterprises are going to want more pre-integrated and pre-tested solutions – infrastructure management software suites – to drive out cost and complexity.

Other VC or angel funded startups tackling various elements of the IT automation – FineGround Networks, nLayers, Opalis Software, OpTier, Paremus, Singlestep Technologies, Symbium, Tsunami Research, and Vieo – are also budding acquisition candidates. “If any of these companies gets traction, it will likely become part of a larger vendor’s broader management stack,” says William Fellows, principal analyst at The 451 Group.

Opsware: Become a big fish or fish wrap

Opsware is one of the few pure-play, public companies dealing with data-center automation. Marc Andreessen, chairman of Opsware and former Netscape wunderkind, believes that an independent software company, without a hardware agenda, is a necessity in the marketplace.

The Opsware software was developed when the company, previously named Loudcloud, provided managed hosting services. That business didn’t work out (the public company was never profitable and was caught in the downturn) and was sold
in August 2002 to Electronic Data Systems. The company refocused its efforts on the data-center automation software that powered Loudcloud’s data centers.

At the core of Opsware is an “automation platform” that consists of two components. The Software Tree maintains information about applications and operating systems, such as how changes to one application could impact other running applications. The Environment Tree contains data about a specific customer’s data center: hardware, service levels, network infrastructure, applications, asset location and business units. Opsware doesn’t require virtualized resources, but does work with VMware and other virtualization platforms. The knowledge base is used to model and test the impact of proposed changes (modifications, patches and upgrades) on production servers and applications before rollout. The Opsware 4.5 system works across heterogeneous networks and devices, and allows customers to write custom policy automation scripts for provisioning, monitoring, management and control. The company also has templates with pre-defined policies for automating the provisioning of data center resources and applications.

Sharmila Shahani, senior VP of marketing at Opsware, says that Inflow, an IT outsourcing company with 13 data centers in the US, increased servers per system administrator from 50 to 200 using Opsware’s software. It took only two weeks to provision and configure Inflow’s 600 Windows servers – down from more than three months. For EDS, Opsware software manages 50,000 servers across 154 data centers. Applying a server patch on 880 servers took only 2 hours with Opsware instead of 445 hours previously, Shahani says.

Maintaining this unique database of interdependencies is a secret sauce that can lock in customers, but longer-term Andreessen knows that standards will evolve to normalize the definitions of data center operations. To that end, Opsware has been a major contributor to the Data Center Markup Language (see page 20), a proposed open standard for profiling IT infrastructure.

So far, EDS is Opsware’s primary software customer, halfway through a three-year, $54-million contract. EDS just renewed the agreement in August, signing an additional $50-million three-year deal that runs through 2008. For the quarter ended July 31, net revenue totaled $8.6 million, of which EDS accounted for 60 percent. Projected revenue for the year is $35 to $37 million.

OPSWARE INFO

Headquarters: Sunnyvale, CA
Founded: September 1999
Employees: 170
Funding: public company (OPSW); $530 million market cap
Key metric: majority of revenue is from EDS; Opsware 4.5 starts at $1,200 per server
URL: www.opsware.com
The company has about 50 customers, including MetLife and the US Department of Energy. With the recent acquisition of Tangram Enterprise Solutions, a North Carolina-based developer of IT asset management software, the company gained 200 customers, including General Electric, Microsoft, Priceline, Sprint, SunGard, Qwest and Wal-Mart.com.

While Opsware calls itself the leading provider of data center automation software, its position is tenuous: The market is still immature and Opsware competitors (many of whom are its partners, including BEA, HP, IBM and Sun) are beginning to build their own IT automation suites. Opsware and other established stand-alone players such as BladeLogic can claim the advantage of a singular focus over their larger competitors, but that is more wishful thinking than reality over the long term. By the end of the decade, basic configuration management and provisioning will be baked into the products and services that the major infrastructure providers offer. It’s not clear where that leaves Opsware.

Andreessen says he is considering moving laterally into automating network management, but isn’t going to tackle the desktop (dominated by Microsoft) or storage, where companies such as Veritas, Computer Associates and EMC have carved out the IT automation piece.

“One of the interesting questions,” Andreessen says, “is whether companies like Cisco, Sun, and EMC will be in a similar business over time. With the move toward

---

**DATA CENTER MARKUP LANGUAGE**

As part of its effort to drive IT automation, Opsware, along with Computer Associates and EDS, is promoting a new XML format, the Data Center Markup Language (DCML).

DCML provides an inventory of data center components and descriptions of the working relationships among them, says Opsware’s Andreessen. DCML could be used to provision or clone a server or even a data center. The founding group has submitted its proposed standard to OASIS, which will take on development of the specification.

A specific XML-based DCML file could include configurations and profiles for servers (patch levels, I/O settings, installed applications, etc.), applications, networking (network devices, firmware versions and protocols), security (firewalls, anti-virus and intrusion detection devices), storage (storage systems and space allocation) and other components of a data center. Different vendors would use this common language so that customers or other applications could manage their products easily.

“DCML is a good idea but it has a long way to go,” says Ferengul. “It’s another semantic standard that vendors have to agree on. Users have to be more activist for agreement on standards to occur. It has a 50-50 chance of succeeding.”

The DMTF’s Common Information Model (CIM) provides a vendor-neutral schema for describing management information, but DCML takes it a step further by integrating the functional relationships among components. It should result in a smoother path to IT automation and utility computing. So far, HP, IBM, Microsoft and Sun haven’t attached themselves to DCML. Microsoft is leaning towards its own Systems Definition Model (SEE PAGE 30) and the grid community is working Web Services Distributed Management through OASIS. Multiple approaches will certainly have the effect of slowing down progress. The industry needs to work collectively to come up with a common, extensible standard.

The founding group has submitted its proposed standard to OASIS, which will take on development of the specification.

A specific XML-based DCML file could include configurations and profiles for servers (patch levels, I/O settings, installed applications, etc.), applications, networking (network devices, firmware versions and protocols), security (firewalls, anti-virus and intrusion detection devices), storage (storage systems and space allocation) and other components of a data center. Different vendors would use this common language so that customers or other applications could manage their products easily.

“DCML is a good idea but it has a long way to go,” says Ferengul. “It’s another semantic standard that vendors have to agree on. Users have to be more activist for agreement on standards to occur. It has a 50-50 chance of succeeding.”

The DMTF’s Common Information Model (CIM) provides a vendor-neutral schema for describing management information, but DCML takes it a step further by integrating the functional relationships among components. It should result in a smoother path to IT automation and utility computing. So far, HP, IBM, Microsoft and Sun haven’t attached themselves to DCML. Microsoft is leaning towards its own Systems Definition Model (SEE PAGE 30) and the grid community is working Web Services Distributed Management through OASIS. Multiple approaches will certainly have the effect of slowing down progress. The industry needs to work collectively to come up with a common, extensible standard.
blade servers, with CPUs, storage and applications, it’s possible to be a layer of automation for monitoring compliance on top of all of those systems.”

**UXComm: Bottom-up automation**

Like Opsware’s software, Veritas’ CommandCentral Service, IBM’s Tivoli platform and other IT automation solutions, UXComm’s AutonomIQ platform is geared to detecting, monitoring and controlling hardware infrastructure. “We have our own XML-based scripting language that works from bare [servers] metal up to operating systems, applications and middleware,” says UXComm CEO Mark Sigal. “We take manual processes or scripts and automate them in a procedural fashion, allowing users to create workflows that sequence, schedule or define conditional events and corresponding policies, which can trigger actions.” UXComm’s platform doesn’t provide virtualization or grid support.

UXComm focuses primarily on modular server environments, with Windows- and Linux-based blade and 1U (a unit of measure – 1.75 inches – for the height of a rack-mounted device) servers. Funded with $7 million by Olympic Venture Partners, Foundation Capital and Intel Capital in October of 2003, Beaverton, OR-based UXComm’s expects to ship beta of its AutonomIQ for Dynamic Server Provisioning product by the end of this month and a production release in January.

UXComm takes a bottom-up approach: It deploys agents directly on the hardware being managed, rather than executing policies on a centralized manager as with Opsware. The agents operate locally and perform functions such as provisioning a server based on workflow definitions or generating policies based on events on the hardware. “We push the decision-making down to local levels, which speeds up execution. We expect that [our software] can cut costs of systems management by a third to a half,” says Sigal. In essence, it retrofits the infrastructure hardware with local intelligence, taking a self-managing approach to IT automation. Like Opsware, AutonomIQ has canned sets of workflows and tools can to create new workflows or modify existing ones. AutonomIQ software will range in price from $100 to $500 per managed server – roughly equivalent to 5 to 10 percent of the hardware cost, Sigal says. So far, UXComm has three pilot customers, including investor Intel, Force Computer (a unit of Motorola) and Ontario-based CygComm, an integrator distributing a speech gateway built around Microsoft’s speech server.
**Sychron: Infrastructure orchestration**

Sychron, based in San Mateo, CA, offers another example of IT automation that gets close to the self-managing aspects of autonomic computing. In May the company introduced Infrastructure Orchestration and began pilot projects with a few global financial services firms. Formed by a group of computer scientists from Oxford University, Sychron received $7.8 million in venture funding from Sigma Partners, Dot-EDU Ventures and First European American Ventures in February 2003.

Whereas Opsware focuses on automating configuration management and provisioning in data centers, Sychron’s software provides real-time allocation or reallocation of X-86-based server pools based on fluctuating resource demands from applications. According to Bill McColl, Sychron’s founder and CTO (and head of Oxford University’s Scalable Parallel Computing Research Group), the company’s policy-driven engine orchestrates hardware and software components within mixed environments to work in parallel at optimal service levels.

Products such as IBM’s Tivoli Intelligent Orchestrator provide orchestration (multi-step workflow) as well, but they are more focused on modeling and scripting data-center best practices than on real-time policy-based automation, McColl says.

Sychron has an extensible, general-purpose language and run-time system for specifying and automatically enforcing business goals and policies across data center infrastructure. Sensor agents from Sychron or other vendors constantly monitor IT systems for changes in variables in the environment such as CPU utilization and bandwidth consumption. Changes are assessed against policies expressed as application priorities, business goals and service-level targets. As conditions change, actions – such as load balancing, auto-provisioning, virtualization, clustering and grid scheduling – are automatically invoked to maintain consistency with the policy constraints across the infrastructure.

“Essentially, we are putting down a software fabric that can interface to sensor [resource-management tools] and effector [load-balancing, configuration, etc.] technologies,” says McColl. Sychron’s orchestration engine integrates with numerous third-party infrastructure products.

The challenge/opportunity for Sychron (as well as for its competitors) is to overcome the complexity of the environments that its customers are trying to manage. “Today
we have multiple models for managing infrastructure—such as HP Open View, Jareva [Veritas], and VMware. We can take all of these and define interfaces to make them work together [for policy-based automation], but you still have the core problem of individual models. The next phase is to move the core modeling of all products into a unified setting,” says McColl, who is focusing Sycrhon on Windows and Linux environments. He thinks that Microsoft’s “design for operations” approach (see page 30) has the most promise for unifying the models and simplifying policy-based automation.

Enigmatec: Execution management
Also hailing from the United Kingdom, Enigmatec has developed a policy engine for automating application execution across common server pools. The company’s Execution Management System (EMS) suite of services and tools allows users to define high-level policies that initiate actions to fulfill policies such as service-level agreements.

Enigmatec was founded in 2001 and has garnered $11 million in funding from Amadeus Capital Partners, Pentech Ventures and Intel, according to Duncan Johnston-Watt, principal founder and CTO and formerly managing director for Fixed Income Technology at Instinet, a Reuters subsidiary. The company is moving its headquarters from London to Silicon Valley this year, and hired former Tibco executive and Contivo founder Indra Mohan as CEO in January.

Like UXComm and Sycrhon, Enigmatec’s software has agents on the managed components, allowing for distributed execution of logic. “Instead of all workflow going through a single server, it’s distributed across all available servers,” says Mohan. “You avoid a single point of failure; if half the data center goes down, the agents can reconfigure it.” Enigmatec’s EMS allows policy and workflow changes to be pushed out to production without having to test every combination of configurations. The company has partnerships with infrastructure software providers such as VMware and Sonic Software. “Integrating with third-party solutions is a critical path to scaling the business: You have to be part of overall ecosystem,” says Mohan.

Initial test uses of Enigmatec’s software are for [IT] disaster-recovery orchestration (automating the recovery of components) and resource management (launching third-party provisioning, for example). In a disaster-recovery scenario, Enigmatec’s
solution would help eliminate the need for a dedicated business continuation data center; EMS software could automatically repurpose hardware from another test environment data center to recover a production data center. The company is targeting Wall Street trading firms and has four active pilot projects, Mohan says.

The Holy Grid

The Holy Grail of IT automation is the integration of virtualization, policy-based automation, Web services/ SOA and open standards operating in a grid-enabled environment. It is the top of the pyramid for IT automation. A user chooses an application or task, selects desired service levels, and the rest happens behind the scenes, masking complexity and keeping human intervention to a minimum.

“It should be feasible to have common architectural principles and standard interfaces at every level in the stack, from individual clusters to data centers to interconnected data centers and management environments,” Foster says. “The enabling software should use the same mechanisms, just like most people today use TCP/IP for communications in a cluster, within a data center and between data centers.”

Foster says that merging grid protocols with Web-services protocols is the means to drive the growth of grid-enabled commercial infrastructure and IT automation. In January of this year, Akamai, the Globus Alliance, HP, IBM, Sonic Software and Tibco proposed new Web-services specifications – WS-Resource Framework and WS-Notification – that integrate grid and Web-services protocols. These new specifications will allow developers to use the same protocols to create “grid services,” which are Web services that conform to a set of interfaces and behaviors that define how a client interacts with a grid component that exposes itself as a service.

The Globus Toolkit: An open-source grid service manager

The Globus Alliance is leading the way in developing the enabling management software for the emerging world of grid services. The Argonne National Laboratory, the University of Southern California’s Information Sciences Institute, the University of Chicago, the University of Edinburgh, and the Stockholm-based Center for Parallel Computers are core members of the Alliance, which focuses mainly on developing open-source software for grid-based science and engineering applications. Foster collaborated with Carl Kesselman of the USC’s ISI and Steve Tuecke of Argonne on
the Globus Toolkit, an open-source software suite that includes services and libraries for resource monitoring, discovery and management, as well as security and file management within grids. The Toolkit lacks metering and billing specifications, but they are expected next year from the Globus development effort. In addition, areas such as problem determination and policy management require more work, Foster says. The Toolkit runs on major operating systems, including Windows.

Today, the Globus Toolkit, based on the Open Grid Services Architecture (OGSA) developed by the Global Grid Forum, is used for enabling computing grids such as the Network for Earthquake Engineering and Simulation. Commercial companies, including grid-oriented firms Avaki, DataSynapse, and Platform Computing, as well as HP, IBM, NEC, Oracle and Sun, are developing IT automation products and operations management control software based on the Globus Toolkit.

Sun, for example, recently introduced its N1 Grid for SAP Infrastructure, which creates a common pool of Sun hardware that can be shared among multiple SAP components. The environment constantly adjusts parameters to maintain service levels. Meanwhile, HP plans to grid-enable all of its applications within the next few years.

The work by the Globus Alliance, the Global Grid Forum, the Enterprise Grid Alliance (a consortium of leading vendors and customers), and standards bodies has yielded progress, but the standards are still in flux. Guaranteeing interoperability and security among thousands of interdependent components, as well as applications that demand special configurations, remains a major challenge.

“Mechanically, we understand a fair bit, including defining the basic plumbing and what it means to build a system that behaves in autonomic fashion, and what interfaces and behavior to push to resources, and at what management level,” Foster says. “It’s partly a question of making individual components smart and resilient, and providing the hooks to get access to information at the management level. Politically we need a critical mass to agree on standard approaches. If IBM, Microsoft, HP and Sun are not in agreement, it’s hard to make progress.”

The major vendors have shown more cooperation on standards in the last year. As with Linux, vendors have to compete more on the quality of their distributions and the services they provide. That doesn’t mean they won’t try to push proprietary extensions for the software or to promote standards that support their agendas, but pressure from customers will drive proprietary software higher up in the stack. The world doesn’t need a dozen languages for automating IT infrastructure.
Cassatt: The Google of IT automation?

Cassatt, a San Jose, CA-based start-up, is attacking IT automation with operations management software that leverages the Globus Toolkit. In September 2003 company CEO and founder Bill Coleman began his own Manhattan Project, getting funding — rumored to be in the $50 million range — from Warburg Pincus and acquiring the people and technology of Unlimited Scale, a Linux-based server-clustering technology company chaired by former Cray Research CEO John Rollwagen.

Cassatt was named after a 19th-century family of artists and entrepreneurs. Alexander Cassatt was president of Pennsylvania Railroad and oversaw the building of Penn Station, a hub that connected New York City to the rest of the country and helped revolutionized the delivery of goods. Mary Cassatt was a leading American impressionist. Coleman believes that his company’s breakthrough (which he isn’t ready to reveal to the world) is unique, and will revolutionize the delivery of IT services. At this point we can’t judge those claims, but Coleman and company have a strong track record of successes and plenty of industry experience.

The investors are betting that Coleman’s company can become the Google of infrastructure management software, outgunning the likes of CA, IBM, HP, Sun, the grid software companies and others who share their visions and goals. Google has built a leading Web-based platform and infrastructure for delivering applications and services. Cassatt is developing what it hopes will be a dominant platform for managing and automating grid-based IT infrastructure.

Coleman has hired an A-list of technologists to join his quest: Steve Oberlin, former president of Unlimited Scale and chief architect at Cray Research; Brian Berliner, a co-founder of Allocity, an application-specific storage management company; Mark Forman, former CIO for the US Office of Management and Budget; and Rich Green, former Sun VP of programming tools. In addition, Coleman hired a Colorado Springs-based team of Sun engineers working on remote distributed management.

Cassatt is taking a holistic approach to automating IT. “You can’t treat [IT automation] as provisioning, reliability, data clustering or fail-over problems, or as a quality of service or scale problem — it’s a superset of them all,” Coleman says. “You can look at it as virtualizing the power plant and the distribution of electrical power.”
The company isn’t going to fully unveil its secrets until the end of year, but Coleman gives hints, such as “converging grid computing and service-oriented architectures.” He also describes Cassatt’s products as a run-time management and execution layer that efficiently scales commodity hardware using industry-standard operating systems. The company plans to start slowly, with the first iteration delivering support for X-86-based hardware, Linux and the Globus Toolkit, initially targeting departments and business units rather than entire enterprises.

Coleman also claims that Cassatt has solved the problem of scaling infrastructure and reducing headcount without degrading performance: “Adding processors [to a cluster] doesn’t scale linearly, so you have to find a new way to scale linearly no matter how large the system. That’s the biggest problem, and we have solved it. I can’t tell you how, but we have over 3,000 processors with 3-percent overhead never going up. You also have to understand what’s important to virtualize. If you try to virtualize memory, for example, you get a declining return for scaling across the network.”

While Coleman is coy about how the company scales computing resources efficiently, he cites a three-dimensional matrix for managing trade-offs and resource allocation:

- Physical infrastructure capacity (processing, storage and network)
- Throughput (how much capacity is need to run an ERP payroll application every Wednesday)
- Quality of service (if you want to go from four-nines to five-nines)

“The three dimensions became the breakthrough by which we manage dynamically what is running and where, and how to map I/O processing and virtual LAN operations. It’s a goal-seeking mechanism that applies policies as a basis for running logical entities and mapping them to the virtual pool of resources,” he says.

Coleman asserts that his start-up has advantages over IBM, HP and others who are building IT automation into their existing products: “We have a more platform-independent point of view. We don’t have business models at risk like the big guys with proprietary hardware and software or the need to re-instrument applications or buy new versions of applications.”

Cassatt’s technology concepts, at least, will resonate with customers (including customers from Coleman’s stint at BEA), but it’s ahead of where most CIOs are thinking and spending today. Cassatt won’t be the only freshly minted company chasing the Holy Grid, but it will likely be the best capitalized and among the most talented.
Following in Red Hat’s footsteps
Ian Foster has joined up with industry veteran Rich Miller to form a commercial company built around the Globus Toolkit. Miller is a Silicon Valley veteran who designed and implemented some of the first computer messaging and conferencing systems for the ARPAnet in the early 1970s, founded several companies (IT/telecom consulting firms Rapport Communication and Telematica) and headed up the now-defunct General Magic’s communication technology (Telescript) business unit. Foster will serve as the company’s chief strategist and sit on the board of directors.

The company hasn’t settled on a name yet, but Miller plans to have 20 to 25 people on staff in Silicon Valley and Chicago by the end of this year. So far, the company is self-financed but is closing a seed round this month. “The objective right now is to say low profile for the rest of this year, and move into the spotlight at the end of the first quarter next year,” says Miller.

The company is targeting the 4.0 release of the Toolkit, which is expected early next year, and will focus on cross-domain grid management. “You have to provide a distributed form of control, where everybody gets resources allocated based on policies, without encroaching upon the autonomous management of any single domain itself,” Miller says.

Miller describes the function of his product as “resource assurance” in an on-demand environment. “For distributed systems, it’s a matter of putting in a reservation for resources a minute or second before they are needed, and having the assurance that they will be available for the application.”

The startup also plans to take pages out of the Red Hat playbook. “For the Globus Toolkit to have commercial adoption, it requires 24-by-7 support and software distributions that are clearly designed to work well in a variety of host environments (operating systems, hardware platforms, network interconnects, job schedulers, etc.), integration with back-end systems and development tools for corporate end users of grid-enabled technology,” Miller says. The company’s initial target markets include financial services, insurance firms and life-science organizations. Miller also says the applications built with the Globus Toolkit could be used to manage enormous data sets such as RFID information. “A data grid [as opposed to a computational grid] makes a great deal of sense for manufacturing and supply chain applications,” Miller says. A data grid uses multiple storage systems from several owners, dividing data across the resources to host a single, common data collection.
Besides Cassatt and Miller’s new companies, IBM, HP and others will compete to deliver grid-centric operations management solutions. “We expect to succeed primarily by sticking to our knitting,” Miller says. “We are focused on commercial support of Globus, and on building connectors and integration piece parts so that IBM and others can grid-enable their product lines.” If the Globus Toolkit gains enough support, as it seems to be doing even at this early stage, the Red Hat model – in which solution vendors are content with standardized, open-source Linux distributions from a few players – may apply.

While the majority of what Miller’s company is doing will find its way into the open-source corpus, he ruled out the development of proprietary adjuncts to the Globus Toolkit. “Cassatt is taking a similar approach to us inside the data center,” Miller says. “Our company is taking Globus Toolkit into more distributed environments. We could be the glue between Cassatt implementations or other resources that use different approaches. The entry ticket for us is the Globus Toolkit, which exposes [interfaces] to the outside that adhere to standards.”

Miller believes that within the next three to four years, grids will be a commonly accepted approach to building out enterprise infrastructure, and the basis for extragrids and community intergrids (shared services across the public Internet).

He may be overly optimistic about how fast the grid-service approach, including robust operations systems managers, will achieve mainstream enterprise adoption. Corporations are just beginning to pilot Web services integration projects and to use virtualization technology. It’s more likely to be a decade before enterprises move fully into grid services. At that point infrastructure will be more of an undifferentiated commodity: a black box that runs in the background and requires increasingly less human intervention.

As Forrester analyst Fichera points out, organizational friction and operational realities often slow adoption to a greater degree than the technology itself. Customers are more ready for less risky islands of automation than a universe of automation. SAP today runs on old-fashioned client/server systems; shifting to SOA, grid-enabled .Net or Java environments is a decade-long process. On the other hand, the stage is set for rapid growth in IT automation as companies make that transition. And the companies that make the shift will scale more smoothly and grow faster, so that early adopters will gain share both by growing internally and by increasing in number.
Microsoft's Plan: Design for Operations on Its Own Turf

While most companies are dependent on standards-body efforts for automating IT, Microsoft has developed its own unique, bottom-up approach. The company gives a nod to the auto industry for inspiration, but the impetus comes from a critical need to make its software more secure, reliable, and manageable. In the 1980s, automakers in the US needed to improve manufacturing methods in order to stay competitive. They determined that the key was to improve the design process, eventually developing the mantra, “design for manufacturing.”

Microsoft has mutated the phrase slightly and is focusing its IT automation efforts on “design for operations.” Its approach seeks to instrument software with automation capability from the day it is conceived. “We need more cross-pollination, with the people focused on running applications involved in the design and development process,” says David Hamilton, director of Microsoft’s Windows and enterprise management division. “We have to work out how to encode operational behavior into applications at design time, and then follow up through their deployment and full lifecycle.”

“Design for operations” is expressed in Microsoft’s System Definition Model (SDM), which is part of the company’s Dynamic Systems Initiative (DSI), a comprehensive initiative to improve the manageability of Windows and everything it touches. Using XML as an underlying data format, the SDM captures the basic structure and definitions embedded in hardware and software components, including information such as configuration schemas, “health” models and operational policies.

“We believe at design time that we can express the desired state of individual components – such as specifying a particular database and what server it runs on, whether it runs in a cluster, and the desired throughput speed,” Hamilton says. “Other vendors view the heavy lifting happening with management tools. You need management tools to track SLAs [service-level agreements] and to orchestrate workflow. We believe that resource management and partitioning of hardware resources isn’t as important as people resources and the manageability of applications.” In other words, designing applications from the ground up for smooth operation and that meet the specific needs of the users is preferable to depending mostly on management tools overlaid on applications automate IT. It’s a perspective that you would expect from Microsoft.
The SDM format has some similarities to other system management standards proposals, such as Data Center Markup Language (see box, page 20) and ongoing efforts from the DMTF. “We are working closely with the standards bodies to build and evolve the DMTF’s CIM specification to deliver a platform-independent, industry-standard server hardware management architecture across diverse IT environments in the data center,” Hamilton says. “Instead of focusing on the individual hardware components, the SDM takes a distributed system view and describes how all the various components of that IT system come together to function as one complete system and what the relationships are between those various components.”

Hamilton says that while the company will focus on developing products and solutions for the Windows platform, it will work with third-party developers to extend the SDM to non-Windows environments.

The SDM is an ambitious, compelling concept – it sounds like photosynthesis for IT, at least in the Windows world. But without getting broad vendor support and being published as an open standard, it will be too Microsoft-centric. And, as with expert systems, which apply artificial intelligence techniques to capturing the specialized domain expertise of individuals for problems such as diagnosing diseases, there is a problem is making sure the knowledge hard-wired into the system is reliable and accurate. In addition, modeling existing, heterogeneous systems is difficult because a substantial amount of legacy and custom code is poorly documented.

According to Hamilton, the company’s quest to move intelligence from management software into the resources themselves will pass through several stages. The first stage is pushing all information into systems management software – the current state of the industry. “In this environment the level of application management is basic; you can manage Exchange through tools that look at the flow of data in and out of the box, but they don’t look inside the application or understand its context,” Hamilton says.

In stage two, vendors provide explicit management information about their applications, which can be used by systems management software, Hamilton says. “For example, the Microsoft Operations Manager (MOM) Management Pack (MP) for Exchange can interpret the behavior and events generated by the application, creating the appropriate alerts, tracking thresholds and delivering context-sensitive reports,” Hamilton says. A net effect of going from stage-1 to stage-2 technology is a significant reduction in the alert-to-ticket ratio, Hamilton notes.
At the highest level, stage 3, most management infrastructure becomes part of the platform. During the initial development process, data about the operation and behavior of an application is collected and stored, and applied according to application-specific management schemas. For example, adding a new application needn’t require any manual configuration; all the information could be embedded in the application, which communicates with the infrastructure for pre-deployment verification, installs itself and operates according to a pre-defined schema.

The major benefit of stage 3 is reduction in the cost and complexity of managing IT is reduced. The journey to stage 3 – full-blown SDM – will take about a decade, Hamilton says. He cites the Management Pack for Microsoft SQL Server as an example IT automation using SDM concepts, though it’s still at stage 2 he says.

During the 10 years it should take SDM to mature and be adopted, Microsoft has and plans to develop an array of management, provisioning and IT automation tools, mostly aimed at Windows environments. The company plans to ship its Microsoft Operations Manager (MOM) 2005 in October, with connectors to other management consoles and more than 50 management packs that monitor applications, model their health and have the capability to correct errors.

Next year, Microsoft will combine MOM 2005 and System Management Server (SMS) 2003, into a new product called System Center.

**Self-assembling IT**

Professor Heylighen doesn’t believe that the notions of policy-automation tools, grid architectures – or IBM’s vision of autonomic computing – are radical enough to fully automate IT. He advocates self-organizing systems, which he defines as functional structures that organize and maintain themselves spontaneously. “This is akin to what Microsoft proposes with its SDM. IBM Research has also developed similar concepts (see page 15).

Applying self-organizing principles to IT automation seems a bit vague and wishful, but dozens of researchers have been studying related concepts such as swarm intelligence, the way insects inhabit self-organized societies, and applying the findings to IT problems. Individual ants deposit pheromones that other ants use to make their foraging decisions, creating an efficient network of paths without a centralized hub.
The collective behavior is orchestrated in a way that appears to be intelligent. Each ant contributes to what the colony as a whole does, so no single ant is in control.

Algorithms to route network traffic more intelligently have been derived from analyzing ant colonies; the most efficient routes are promoted and the less efficient paths are abandoned dynamically based on feedback loops among the nodes.

Google’s PageRank system, for example, is also self-organizing. Google automatically interprets links using various inputs to rank pages for display. The contents of the Web are thus organized according to policies that Google establishes through its algorithms. But turning a data center into a self-organizing, autonomic entity is a more challenging problem. Yes, Google has more than 100,000 servers, but given that they are running mostly a single application on a single kind of hardware (mostly on home-grown Linux/Intel boxes), the challenge is mitigated. Now, however, Google has added complexity: back-office systems, ad serving, billing, Gmail, Froogle, Orkut, IPO auctions, regulatory compliance, and other applications. The company will be a perfect candidate for the grid services architecture and for IT automation.

IBM Research has an autonomic computing concept similar to Heylighen’s self-organizing IT that it describes as goal-driven self-assembly, a decentralized autonomic computing framework based on multiple interacting agents. Each component, ranging from databases and servers to workload managers and policy repositories, is responsible for managing itself and what it controls. In goal-driven self-assembly, each component knows what role to play in the system and its goals expressed as policies, such as “I am a database and must perform optimally at 1000 transactions-per-second.”

When a component initializes, it contacts a registry of other self-managing components that can offer the requisite services to build an application. For example, an application server will go to the registry to find a database that agrees with its goals, entering into a kind of performance contract. Then the database will look for an appropriate storage system to fulfill its needs. When all the pieces are assembled, they are dynamically bound together. If any of the components fails, a system management function might solve the problem by checking a directory (like a UDDI Web service directory) substitute component that works.
IBM’s White says that the difficulty with goal-seeking self-assembly is in determining how to state policies in ways that are actionable and how to translate them from the invoking component to other parts of the system that have to understand the communication. It all goes back to establishing standards for the interactions and for the components.

Heylighen believes that a hard-wired, policy-based approach with standard interfaces isn’t sufficient to foster autonomic computing on a broad scale. “If you look at human language, it evolved without a priori knowledge of how to communicate. It’s a good strategy to start with a default case [pre-defined semantics, such as ooga ooga! = yes yes!], but there are constantly new functions and interpretations. The Semantic Web, for example, is designed to have a broad, shared ontology, but how do you get everyone to agree on a single ontology and let it evolve organically? Older devices, for example, may not understand the semantics of newer devices. Developers come up with meta-language or meta-agents, but it gets too complicated. “

Self-organization is not a question of technology, according to Heylighen. It’s more a question of the design and the science behind it, and a generalized architecture. “Rather than relying on semantic primitives, agents have to develop their own language and behavior,” says Heylighen. “Similar to a neural network, you have inputs and outputs, but a lot of complexity underneath. Agents form neural networks between each other, and the output agents are what the user sees and monitors.”

A 2001: A Space Odyssey outcome, in which computers and agents take on a life of their own in conflict with the humans who engineered the systems, is unlikely, Heylighen says. “I see it evolving into symbiotic system in which people, devices and networks adapt to each other and form a global brain. All the elements collaborate in an intelligent way, but not any one is in control of all – it’s more like a swarm.”

The global brain sounds seems counter-intuitive and fanciful, given the propensity of humans to dominate one another. As IT infrastructure evolves and becomes more autonomic and adaptive, applications will sprout as needed to achieve goals derived by the system. Devising the universal protocols for interaction that support unrestricted self-organization are probably more on the order of getting to the bottom of string theory in physics. Now we are entering the realm of squishy realm of artificial life and genetic algorithms, which are beyond the scope of this issue.
Resources & Contact Information

Rich Miller, 1 (650) 463-1777; miller@globussoftware.com
William Fellows, The 451 Group, 44 (207) 299-7766; william.fellows@the451group.com
Ian Foster, Argonne National Laboratory, 1 (630) 252-4619; foster@mcs.anl.gov
Bill Coleman, Cassatt, 1 (408) 437-7600; bill.coleman@cassatt.com
Mark Forman, Cassatt, 1 (408) 437-7600; mark.forman@cassatt.com
Don LeClair, Computer Associates, 1 (631) 342-6000; donald.leclair@ca.com
Duncan Johnston-Watt, Enigmatec, 1 (212) 837-7928; duncan@enigmatec.com
Indra Mohan, Enigmatec, 44 (20) 739-8202; indra@enigmatec.com
Richard Fichera, Forrester Research, 1 (617) 613-6089; rfichera@forrester.com
Francis Heylighen, Free University of Brussels, 32 (2) 644-2677; fheyligh@vub.ac.be
Nick van der Zweep, Hewlett-Packard, 1 (650) 857-1501; nick.vanderzweep@hp.com
Alan Ganek, IBM, 1 (914) 766-1850; fax, 1 (914) 766-1638; ganek@us.ibm.com
Steve White, IBM, 1 (914) 784-7368; fax, 1 (914) 784-6912; srwhite@watson.ibm.com
Robert Le Blanc, IBM, 1 (512) 286-2100; fax, 1 (919) 254-7211; rjl@us.ibm.com
Curt Cotner, IBM Fellow, 1 (408) 463-2747; fax, 1 (408) 463-2614; cotner@us.ibm.com
Prasad L. Rampalli, Intel, 1 (916) 356-2694; prasad.rampalli@intel.com
Ravi Subramaniam, Intel, 1 (408) 765-3566; ravi.subramaniam@intel.com
David Nelson-Gal, Interwoven, 1 (408) 220-7435; dgal@interwoven.com
Scott Davis, Katana Technology, 1 (978) 849-1210; fax, 1 (978) 849-1299; scott@katana-technology.com
Cary Feren gul, META Group, 1 (630) 588-0988; corey.ferengul@metagroup.com
David Hamilton, Microsoft, 1 (425) 703-5577; dhamilton@microsoft.com
Marc Andreessen, Opsware, 1 (408) 744-7507; pmarca@opsware.com
Sharmila Shahani, Opsware, 1 (408) 212-5292; sshahani@opsware.com
Bill McColl, Synchron, 1 (650) 931-0747; mccoll@synchron.com
Mark Sigal, UXComm, 1 (503) 578-2890; fax, 1 (650) 578-2889; msigal@uxcomm.com
Diane Greene; VMware, 1 (650) 475-5000; diane@vmware.com
Mendel Rosenblum, VMware/Stanford University, 1 (650) 384-4856; mendel@vmware.com

For further reading:
“Autonomic Computing: IBM’s Perspective on the State of Information Technology,” by Paul Horn:
The Argonne National Laboratory: http://www.anl.gov
University of Southern California’s Information Sciences Institute: http://www.isi.edu
Center for Parallel Computers: http://www.pdc.kth.se
Release 1.0 Subscription Form

Complete this form and join the other industry executives who regularly rely on Release 1.0 to stay ahead of the headlines. Or if you wish, you can also subscribe online at www.release1-0.com.

Your annual Release 1.0 subscription costs $795 per year ($850 outside the US, Canada and Mexico), and includes both the print and electronic versions of 11 monthly issues; 25% off the cover price when you order from our online archives; a Release 1.0 binder; the bound transcript of this year’s PC Forum (a $300 value) and an invitation to next year’s PC Forum.

NAME ________________________________________________
TITLE ________________________________ COMPANY _______________________
ADDRESS _____________________________________________________________
CITY ____________________ STATE ______ ZIP ________ COUNTRY __________
TELEPHONE ___________________________ FAX ______________________
E-MAIL* ____________________________ URL ______________________

*Personal e-mail address required for electronic access.

☐ My colleagues should read Release 1.0, too!
Send me information about multiple copy subscriptions and electronic site licenses.

☐ Check enclosed ☐ Charge my (circle one): AMERICAN EXPRESS MASTER CARD VISA
CARD NUMBER ________________________________ EXPIRATION DATE __________
NAME AND BILLING ADDRESS ________________________________________________
SIGNATURE ______________________________________________________________

Please fax this form to Brodie Crawford at 1 (212) 924-0240.

Payment must be included with this form. Your satisfaction is guaranteed or your money back.

If you wish to pay by check, please mail this form with payment to: EDventure Holdings, 104 Fifth Avenue, 20th Floor, New York, NY 10011, USA. If you have any questions, please call us at 1 (212) 924-8800; e-mail us@edventure.com; www.release1-0.com.

Get our RSS feeds: We now have RSS feeds for Release 1.0 and Esther’s weblog.
Subscribe to them at http://www.edventure.com/rss