All Systems Down

By Scott Berinato (11 April, 2003)

https://www.cio.com.au/article/65115/all_systems_down/

(Archived)

A blow-by-blow record of one of the worst health-care IT crises in history and what CareGroup CIO John Halamka learned from it.

Reader ROI

- What can happen if an old network is asked to carry new applications
- How standard disaster recovery protocols can fail
- Lessons learned from the four-day crisis

Among the 30-odd CIOs who serve Boston's world-famous health-care institutions, John Halamka is a star among stars. He has been CIO of the caregroup health organisation and its premier teaching hospital — the prestigious Beth Israel Deaconess Medical Centre — since 1998. He helps set the agenda for the Massachusetts Health Data Consortium, a confederation of executives that determines health-care data policies for New England. Until 2001, the 40-year-old Halamka also worked as an emergency room physician, but he gave that up to take on the additional responsibilities of being CIO of Harvard Medical School in 2002. However, as a globally recognised expert on mushroom and wild plant poisonings, he is still called when someone ingests toxic flora.

All of this has earned Halamka a considerable measure of renown. For two years running, InformationWeek in the US named Halamka's IT organisation number one among hospitals in its yearly ranking of innovative IT groups. In September 2002, CareGroup was ranked 16th on InformationWeek's list of 500.

Two months later, Beth Israel Deaconess experienced one of the worst health-care IT disasters ever. Over four days, Halamka's network crashed repeatedly, forcing the hospital to revert to the paper patient-records system that it had abandoned years ago. Lab reports that doctors normally had in hand within 45 minutes took as long as five hours to process. The emergency department diverted traffic for hours during the course of two days. Ultimately, the hospital's network would have to be completely overhauled.

This crisis struck just as health-care CIOs in the US are extending their responsibilities to clinical care. Until recently, only ancillary systems like payroll and insurance had been in the purview of the CIO. But now, in part because of Halamka and his peers, networked systems such as computerised prescription order entry, electronic medical records, lab reports and even Web conferencing for surgery have entered the life of the modern hospital. These new applications were something for health-care CIOs to boast about, and Halamka often did, even as the network that supported the applications was being taken for granted.

“Everything’s the Web,” Halamka says now. “If you don’t have the Web, you’re down.” Until last November 13, no one, not even Halamka, knew what it really meant to be down. Now, in the wake of the storm, the CIO is calling it his moral obligation to share what he’s learned.

“I made a mistake,” he says. “And the way I can fix that is to tell everybody what happened so they can avoid this.” Sitting in his office three weeks after the crash, Halamka appears relaxed and self-possessed. There’s another reason he’s opening up, talking now about the worst few days of his professional life at CareGroup. “It’s therapeutic for me,” he says, and then he begins reliving the disaster.

Wednesday - THE NETWORK FLAPS

On November 13, 2002, a foggy, rainy Wednesday, Halamka was alone in his office at Beth Israel when he noticed the network acting sluggishly. It was taking five or 10 seconds to send and receive e-mail. Around 1.45pm, he strolled over to the network team to find out what was up. A few of his 250 IT staff members, who range from low-level administrators to senior application developers, had already noted the problem. They told him not to worry. There was a CPU spike — a sudden surge in traffic. RCA, one of the core network switches, was getting pummelled. From where, they didn’t know. It might have to do with a consultant who was working on RCA, preparing it for a network remediation project.

“We happened to have had a guy in there,” recalls Russell Rusch of Callisma, the company leading the remediation project. “We knew [the hospital] had had similar incidents in the past few months.” Those previous CPU spikes lasted anywhere from 15 minutes to two hours, he says. Then they worked themselves out. Like indigestion.

Halamka’s team decided to begin shutting down virtual LANs, or VLANs. They would turn off switches to isolate the source of the problem, much in the same way one would go around a house shutting off lights to find out which one was buzzing. Halamka thought the plan sounded reasonable. It was a mistake.

Shutting switches forced other switches to recalculate their traffic patterns. These calculations were so complex that those switches gave up doing everything else.

Traffic stopped. The network was down. Within 15 minutes, by 2pm, the team reversed course and turned all the switches back on. A sluggish network, they figured, was preferable to a dead one.

For the rest of the day and into the night, the network flapped — a term Halamka uses to describe the network’s state of lethargy dotted by moments of availability and, more often, spurts of dead nothing. The team searched for the cause. Around 6 o’clock, when most of the doctors, nurses, staff and students left, the network settled down. Finally, at 9pm, the IT staff found its gremlin: a spanning tree protocol loop.

Spanning tree protocol is like a traffic cop. Data arrives at a switch and asks spanning tree for directions. Say, from John’s server to Mary’s desktop. Spanning tree calculates the shortest route. It then blocks off every other possible route so that the data will go straight to its destination without having to make decisions at other crossroads along the way.

But spanning tree will look only as far out as seven intersections. Should data reach an eighth intersection, called a hop in networking, it will lose its way. Often, it will drive itself into a loop. This clogs the network in two ways. First, the looped traffic itself gums up the works. Then, other switches start to use their computing horsepower to recalculate their spanning trees — to make up for the switch that is directing traffic in a loop — instead of directing their own traffic.

That's what happened at Beth Israel Deaconess. On Wednesday, a researcher uploaded data into a medical file-sharing application, and it looped. The data was several gigabytes, so it clogged the pipes. Then, when Halamka's team turned off a switch at 1.45pm, it was as if one cop closed an intersection and every other cop stopped traffic in all directions to figure out alternate routes.

Halamka's team now knew what happened, if not where it happened. Standard troubleshooting protocol for spanning tree loops calls for cutting off redundant links on the network. "What you're doing is eliminating potential spots where there are too many hops, and creating one path from every source to every destination," Callisma's Rusch says. "It might make for a slower environment" — without backup — "but it should make for a stable environment."

"We cut the links," Halamka says. "It seemed to work. We went home feeling great. We had figured it out."

Thursday - CLOGGED ARTERIES

Hospitals come alive early. By 7am, doctors and nurses started to send some of Beth Israel Deaconess's 100,000 daily e-mails. The pharmacy began filling prescriptions, transferring the first bits of the 40 terabytes that traverse the network daily. Some of the 3000 daily lab reports were beginning to move. By 8am, the network again started acting as if it were flying into a headwind. Halamka realised the network had settled down the night before only because hardly anyone was using it. When the workday began in earnest, CPU usage spiked. The network started flapping. The problem hadn't been fixed.

Halamka's team scrambled to find other possible sources of the trouble. One suspect was CareGroup's network of outlying hospitals in Cambridge, Needham, Ayer and elsewhere in Massachusetts. They operated as a distinct network that plugged into Beth Israel Deaconess. The community hospitals' network was sluggish, and a billing application wasn't working, according to Jeanette Clough, CEO of Mount Auburn Hospital in Cambridge, which serves as the hub for the outlying hospitals' network. The easiest thing to do would be to cut the links, eliminating the potential for spanning tree loops. But that would isolate the outlying hospitals. Instead, the IS team, along with Callisma engineers, chose a more complex option. They would try converting from switching to routing between the core network and the outlying hospitals. That would eliminate spanning tree issues while keeping those hospitals connected. They tried for seven hours, and, for arcane reasons that have to do with VLAN Trunking Protocol (VTP), they never got the routing to work. The network flapped all day. Around midmorning, as Halamka was explaining the routing strategy to CareGroup executives in an ad hoc meeting, a patient, an alcoholic in her 50s, was admitted to Beth Israel Deaconess's ICU. Dr. Daniel Sands, a primary care physician and director of the hospital's clinical computing staff, saw her. She had what Sands calls "astounding electrolyte deficiencies", a problem common to people who drink their meals. In fact, Sands says, "It was incredible she was alive.

“I needed to be careful with this woman. I needed to try treatments based on lab reports and then monitor progress and adjust as I went,” recalls Sands. “But all of a sudden, we couldn’t operate like that. Usually I get labs back in less than an hour; they were taking five hours, and here I have a patient who could die. I was scared.” (The patient would survive.)

At 4pm, Halamka met with a minicrisis team that included the head of nursing, the heads of the lab and the pharmacy, and hospital COO Dr Michael Epstein. “Even then,” Halamka says, “I’m still saying: ‘We’re one configuration change away’, and my assumption is things will be back up soon.”

But his team was tense and frustrated. CareGroup’s help desk had been flooded with calls. They were hearing everything from “I can’t check my e-mail” to “I don’t know if the blood work I just requested went through”.

At 3.50pm, Beth Israel closed its emergency room. It stayed closed for four hours, until 7.50pm, according to Massachusetts Department of Public Health documents.

It was at the 4pm meeting that COO Epstein says he realised “this was more than a garden-variety down-and-up network”. Clinical users, like Sands, were signalling that they were worried. Epstein and Halamka, along with hospital executives and network consultants, decided to take extreme measures. They called Cisco Systems, the hospital’s equipment and support vendor. Cisco responded by triggering its Customer Assurance Program (CAP), a bland name that belies how rare and how serious CAPs are. CAP means Cisco commits any amount of money and every resource available until a crisis is resolved.

CAP was declared shortly after 4pm. By 6pm, a local CAP team from nearby Chelmsford, Massachusetts, had set up a command centre at the hospital and initiated “follow the sun” support — meaning additional staff at Cisco’s technical assistance centres would be plugged in to the crisis until their workday ended, when they’d hand off support to a similar group a few time zones behind them.

First, the CAP team wanted an instant network audit to locate CareGroup’s spanning tree loop. The team needed to examine 25,000 ports on the network. Normally, this is done by querying the ports. But the network was so listless, queries wouldn’t go through.

As a workaround, they decided to dial in to the core switches by modem. All hands went searching for modems, and they found some old US Robotics 28.8Kbps models buried in a closet. Like musty yearbooks pulled from an attic, they blew the dust off them. They ran them to the core switches around Boston’s Longwood medical area and plugged them in. CAP was in business.

By 9pm, they had pinpointed the problematic spanning tree loop. The Picture Archive Communication System (PACS) network, for sharing high-bandwidth visual files and other clinical data, was 10 hops away from the closest core network switch, three too many for spanning tree to handle.

And that’s when the dimensions of the problem fully dawned on the team members: They were struggling with an outmoded network. In September 2002, Halamka had hired Callisma’s Rusch to audit CareGroup’s infrastructure. When Rusch finished, he told Halamka: “You have a state-of-the-art network — for 1996.”

Halamka’s network was all Layer 2 switches with no Layer 3 routing. Switching is fast, inexpensive and relatively dumb, and it relies on spanning tree protocol. Routing is more expensive but smarter. Routers have quality-of-service throttles to control bandwidth and to isolate heavy traffic before it overwhelms

the network. State-of-the-art networks in 2002 have routing at their core. In 1996, CareGroup's network was Beth Israel Hospital, and at its core was a switch called Libby030. In October of that year, the hospital merged with Deaconess Hospital. Deaconess's network was plugged into Libby030. Other systems were tacked on in the same way. In 1998, CareGroup connected PACS to what used to be Deaconess Hospital. A year later, CareGroup linked a new data centre and its two core switches (RCA and RCB) to Libby030. There would be a fourth core switch added and a skein of redundant links, but Libby030 remained the main outlet. Halamka now understands that this was a "network of extension cords to extension cords. It was very fragile," he says.

To fix the problem, the CAP team decided to put a Cisco 6509 router between the core network and PACS, eliminating spanning tree protocol and its seven-hop limitation. (The 6509 also has switching capabilities, so the team decided to kill three switches inside PACS and use the 6509 for that too.)

Soon after 9pm, a 747 with a Cisco 6509 on board left Mineta International Airport in San Jose bound for Boston's Logan International Airport. The local CAP team spent the night rebuilding the PACS network, a feat Halamka talks about with a fair bit of awe: The first time around, PACS took six months to build. After working through the night, the team was momentarily disheartened on Friday morning to see that, despite PACS being routed, the network was still saturated. But they rebooted Libby030 and another core switch, which brought out the smiles. "We rebooted and things looked pretty," Halamka says.

Friday - BACK TO PAPER

By 8AM, the network started to flap again. At 10am, Halamka and COO Epstein decided to shut down the network and run the hospital on paper. The decision turned out to be liberating. "We needed to stop bothering the devil out of the IT team," says Epstein. Shutting down the network also freed Sands and the hospital's clinicians. Some had already given up on the computers but felt guilty about it. But "once the declaration came that we were shutting down the network, we felt absolved of our guilt", Sands recalls.

The first job in adapting to paper is to find it: prescription forms, lab request forms. They had been tucked away and forgotten. And many of the newer interns had never used them before. On Friday, they were taught how to write prescriptions. When Sands had to write one, it was his first in 10 years at CareGroup. "When I do this on computer, it checks for allergy complications and makes sure I prescribe the correct dosage and refill period. It prints out educational materials for the patient. I remember being scared. Forcing myself to write slowly and legibly." At noon, Epstein came in to lend a hand . . . and walked into 1978. Epstein worked the copier, then sorted a three-inch stack of microbiology reports and handed them to runners who took them to patients' rooms where they were left for doctors. (There were about 450 patients at the hospital.)

In time, the chaos gave way to a loosely defined routine, which was slower than normal and far more harried. The pre-IT generation, Sands says, adapted quickly. For the IT generation, himself included, it was an unnerving transition. He was reminded of a short story by the Victorian author E M Forster, "The Machine Stops", about a world that depends upon an über-computer to sustain human life. Eventually, those who designed the computer die and no one is left who knows how it works.

“We depend upon the network, but we also take it for granted,” Sands says. “It’s a credit to [Halamka] that we operate with a mind-set that the computers never go down. And that we put more and more critical demands on the systems. Then there’s a disaster. And you turn around and say: Oh my God.”

Halamka had become an ad hoc communications officer for anyone looking for information. Halamka was the hub of a wheel with spokes coming in to him from everywhere — the CAP team, executive staff, clinicians and the outlying hospitals. Halamka leaned on his emergency room training at the Harbor-UCLA Medical Centre in Los Angeles, during the height of gang violence in the 90s. Rule one: stay calm and friendly.

“But I’ll be honest, 48 hours into this, with no sleep, the network’s still flapping, I had a brave face on, but I was feeling the effects,” Halamka recalls. “I was feeling the limitations of being a human being. You need sleep, downtime. You need to think about shifts, or humans will despair.” He found himself dealing with logistics that had never occurred to him: Where do we get beds for a 100-person crisis team? How do we feed everyone? He improvised.

“You don’t know the details you’re missing in your disaster recovery plan until you’re dealing with a disaster,” he says. For example, the paper plan was, in essence, the Y2K plan. Besides the fact that it was dated, it didn’t address this kind of disaster. Recovery plans are usually centred on lost data or having backups for lost data, or the integrity of data. At Beth Israel Deaconess, the data was intact. It was just inaccessible.

That led to Halamka’s chief revelation: You can’t treat your network like a utility.

“I was focusing on the data centre. And storage growth. After 9/11, it was backup and continuance. We took the plumbing for granted. We manage the life cycle of PCs, but who thinks about the life cycle of a switch?”

This is a valuable insight. Networks indeed have got less attention than applications. But at the same time, Callisma’s Rusch says, he hadn’t seen a network as archaic as Beth Israel’s in several years. “Many have already got away from that 1996 all-switched model,” he says. “There are probably a couple of others like this out there.”

Others agree with Rusch’s assessment. “I think the danger is people start thinking the whole health-care IT industry is flawed and a train wreck waiting to happen,” says the CIO of another hospital. “It’s not. We all watched the heroic effort they made over there, but we’re not standing around the water cooler talking about how nervous we are this will happen to us. We’ve had these issues. They scared us enough a few years ago that we took care of the architecture problem.”

Halamka retreated to his office late Friday night. He lay down on the floor, pager in one hand, mobile phone in the other, and fell asleep for the first time in two days. Two hours later, his mobile phone rang.

Saturday - HELPLESSLY HOPING

Half awake, Halamka heard a staffer tell him they had found two more spanning tree errors, one at a facility called Research North and one in cardiology. Both had eight hops, one too many. They planned to cut the redundant links and move the traffic to the core network.

No one knew for sure how severely this would tax poor Libby030 and its counterparts. The team decided to build a redundant core with routing infrastructure as a contingency plan that would bring CareGroup out of 1996 and into 2002 in terms of its network.

At 8am, two more Cisco 6509 routers (with switching capabilities) arrived from San Jose. Three hours before that, a trio of Cisco engineers from Raleigh, North Carolina, landed in Boston. They spent all day building a redundant network core.

Sands felt uncomfortable doing rounds that morning. “Patients sort of expect you to know their histories,” he says. “But without that dashboard of information I’d get from the computer, I had to walk up to patients I had treated before and ask basic questions like, What allergies do you have? Even if I thought I remembered, I didn’t trust my memory. It was embarrassing, and I was worried.”

Progress on the network was slow. No one wanted to suggest that the current tack — building the redundant network core while severing the redundant links — was definitely the answer. At 9am, Halamka met with senior management, including CareGroup CEO Paul Levy. “I can’t tell you when we’ll be functioning again,” Halamka confessed. Admitting helplessness is not part of Halamka’s profile. “You never catch John saying: I’m scared, or I messed up,” says one of his peers from the Health Data Consortium. “This had to be hard for him.”

“When John told us he couldn’t tell us when we’d be up, we stopped having him as part of our twice-a-day reports,” Epstein recalls. The intent was to free Halamka from his communications duties so that he could focus on the problem. But Epstein was also becoming frustrated. He recalls thinking that “we didn’t want to keep sending out memos to the staff that said: Just kidding, we’re still down.”

“If I had felt, in the heat of the battle, that someone could have done a better job than me, if I felt like I was a lesion, then I would have stepped aside,” Halamka says. “At no time did I think this, and at no time was I fearful for my job. Am I personally accountable for this whole episode? Sure. Absolutely. Does that cause emotional despair? Sure. But I had to fix it.” Saturday night, with the redundant core in place, Halamka turned on the network. It hummed. There was clapping and cheering and backslapping among the team, which had grown to 100. Halamka passed around bottles of Domain Chandon champagne that his wife had bought at Costco.

Then he went home. At 1am, his pager woke him.

Another CPU spike.

Sunday - AND ON THE FIFTH DAY, HALAMKA RESTED

The problem was simple: a bad network card in RCB, one of the core switches. They replaced the card. Halamka went back to sleep. Beep. 6am. This time, it was a memory leak in one of the core switches. The CAP team quickly determined the cause: buggy firmware, an arcane VLAN configuration issue. They fixed it.

All day, the team documented changes. Halamka refused to say the network was back, even though it was performing well. “Let us not trust anyone’s opinion on this,” he recalls thinking. “Let us trust the network to tell us it’s fine by going 24 hours without a CPU spike.”

Monday - BACK IN BUSINESS

Halamka arrived at his office at 4am, nervous. He launched an application that let him watch the CPU load on the network. It reads like a seismograph. Steep, spiky lines are bad, and the closer together they are, the nastier the congestion. At one point on Thursday, the network had been so burdened that the lines had congealed into thick bars. Around 7.30am, as the hospital swung into gear, Halamka stared at the graph, half expecting to see the steep, spiky lines.

They never came. At noon, Halamka declared “business as usual”. The crisis was over. It ended without fanfare, Halamka alone in his office.

The same way it had started.

Taking Action

Beth Israel Deaconess CIO John Halamka learned two critical lessons from his four-day disaster.

LESSON 1

Treat the network as a utility at your own peril.

ACTIONS TAKEN:

- Retire legacy network gear faster and create overall life cycle management for networking gear.
- Demand review and testing of network changes before implementing.
- Document all changes, including keeping up-to-date physical and logical network diagrams.
- Make network changes only between 2am and 5am on weekends.

LESSON 2

A disaster plan never addresses all the details of a disaster.

ACTIONS TAKEN:

- Plan team logistics such as eating and sleeping arrangements as well as shift assignments.
- Communicate realistically — even well-intentioned optimism can lead to frustration in a crisis.
- Prepare baseline, “if all else fails”backup, such as modems to query a network and a paper plan.
- Focus disaster plans on the network, not just on the integrity of data.