HARVARD RESEARCH GROUP,Inc.

1740 MASSACHUSETTS AVENUE • BOXBOROUGH, MASSACHUSETTS 01719 • Tel (978) 263-3399

Vol1 The High-Availability Challenge 1999

Recently Harvard Research Group (HRG) completed the analysis of a survey of users of high availability servers. HRG's Primary Research Group interviewed more than 200 self-defined users of high availability servers from the Fortune 2000.

Questions answered included the following:

1. Do vendor offerings meet user requirements?
2. Do users understand vendor claims and terminology?
3. Are users getting what they want?
4. Are the HRG Availability Environments useful in differentiating vendor claims?

When we asked the respondents to define high availability. Almost half the respondents defined HA as -- 7 X 24 operation.

Survey responses came from the following Vertical Industries: Banking, Insurance, Utilities, Manufacturing, Telecom, Transportation, with all others contributing 8.8% of the responses.

All respondents identified themselves as users of highly available servers with first-hand knowledge of the operation of their servers.

Availability Environments Defined: HRG has defined availability in terms of the impact a system being unavailable to perform work has on the activity of the business and consumer (end user) of the service, rather than the technologies used to achieve it. The five Availability Environments (AE) below define availability in terms of the impact on the both the business and the end user or consumer:

- **Fault Tolerant (AE-4) –** Business functions that demand continuous computing and where any failure is transparent to the user. This means no interruption of work; no transactions lost; no degradation in performance; and continuous 24x7 operation.
- **Fault Resilient (AE-3)** – Business functions that require uninterrupted computing services, either during essential time periods, or during most hours of the day and most days of the week throughout the year. This means that the user stays on-line. However, current transaction may need restarting and users may experience performance degradation.
- **High Availability (AE-2)** - Business functions that allow minimally interrupted computing services, either during essential time periods, or during most hours of the day and most days of the week throughout the year. This means user will be interrupted but can quickly relog on. However, they may have to rerun some transactions from journaled files and they may experience performance degradation.
- **Highly Reliable (AE-1)** – Business functions that can be interrupted as long as the integrity of the data is insured. To the user work stops and uncontrolled shutdown occurs. However, data integrity is ensured.
- **Conventional (AE-0)** – Business functions that can be interrupted and where the integrity of the data is not essential. To the user work stops and uncontrolled shutdown occurs. Data may be lost or corrupted.

Disaster Recovery is applicable to any of the HRG Availability Environment Definitions (AEs). Disaster Recovery provides for remote backup of the information system and keeps it safe from the imact of disasters such as an earthquake fire, flood, hurricane, power failure, vandalism, or an act of terrorism.

HRG is actively working with vendors and users to characterize availability in terms of additional dimensions, including reliability the likelihood of an outage and reparability The Parameters of Availability

HA Servers Installed at Sites Interviewed

69 HA servers at AE4
83 HA servers at AE3
79 HA servers at AE2
25 HA servers at AE1

256 HA Servers Total

The vendor with the most HA systems installed at the sites surveyed (*greater than 1/3*) was IBM

Several of the sites with Compaq servers said these servers were run mirrored with Novell's SFT III or Vinca providing the capability of mirroring one server on another and, in case of failure, being able to rapidly switch servers.

Tandem and Stratus servers are normally considered to be fault-tolerant (Availability Environment 4). However, in some cases, respondents also categorized these servers as Availability Environment 3 and Availability Environment 2. One reason may be that these servers were not operationally managed to achieve AE 4 performance.

Operating Systems in Use For high availability servers, at the sites surveyed, the percentage of sites running proprietary operating systems is significant. 60% of the sites run a proprietary operating system, 19% run UNIX and less than 6% run WNT. IBM servers have the largest representation of any vendor's servers in the survey. Most of the IBM servers run a proprietary operating system. 75% of the IBM servers run one of the S/390 operating systems and 14 % run OS/400. The following table examines the distribution of servers by Availability Environment, server type (high-end, mid-range, LAN server), and operating system – proprietary, UNIX, WNT, network operating systems (NOS).

Operating Systems in use

Server type/ OS	AE 1	AE 2	AE 3	AE 4	Total
High-end (proprietary)	6	29	33	18	66 (33%)
mid-range (proprietary)	8	13	14	20	55 (28%)
mid-range (UNIX)	4	18	10	7	39 (19%)
LAN server (WNT OS2)	3	6	6	4	19 (9%)
LAN server (NOS)	3	5	7	8	23 (11%)

Satisfaction with HA Vendors Users were asked to rate their high availability system vendor on: price, total system cost, service & support, ability to recover from failures, ease of use, upgrade path, openness, ease of configuration/customization, and availability of applications.

Satisfaction Ratings

Attribute	Average Rating
price	6.8
total cost	6.9
service and support	7.7
ability to recover	7.9
ease of use	7.5
upgrade path	7.6
openness	6.6
ease of configuration/customization	6.9
availability of applications	7.7

The table above shows the average rating for each attribute. Ratings are based on a scale of 1 - 10 where 1 = completely dissatisfied and 10 = completely satisfied.

Vendor Mindshare HRG asked the respondents to list the top 5 high availability vendors. 92% identified one vendor; 82% identified two vendors; 71% identified three vendors with only 55% identifying four vendors and an even smaller 20% identifying five vendors. In all cases the installed vendor was mentioned as one of the top three high availability vendors.

Enabling Technologies Respondents were asked to identify the primary aspect of their server that made it highly available. Over half the respondents attribute high availability to a software component. Hardware and hardware dependent technologies were used by 35% of the respondents. The following table shows the distribution of servers by HA enabling technology at each Availability Environment

Enabling Technology	AE 1	AE2	AE 3	AE 4	Total
RAID/disk mirroring	5	10	15	6	36
Database/ TP monitor	1	17	16	9	43
Clustering	-	4	4	5	13
Part of OS	1	8	9	5	23
Switching SW (SFT)	2	1	6	4	13
done in application	1	7	2	2	12

Applications HRG asked respondents what business functions or applications were executed on highly available servers. Only the primary function or application was recorded.

Applications Deployed on HA Servers

Application	Frequency	%
Accounting/Finance	36	20
Office, general business	19	11
Database	10	6
Transaction Processing Monitor	9	5
Banking	32	18
Claims Processing	22	12
Manufacturing	17	10
Dispatching/tracking	14	8
Customer service	13	7
Order Entry/processing	5	3

System Age, Upgrade History and Purchasing Plans The median age of servers represented in this study was three years old, with the mean being four years old. There were several systems that in today's terminology are often called "legacy systems" – 24 had been in service for more than a decade – proving that older, stable solutions are holding their value. HRG's study found servers have been upgraded on average 4.23 times. The median upgrade frequency is two.

User Satisfaction (Sat.) (x) System Age

Age of System	# of Servers	% of Total	Ave Sat
<= 1 year	47	23	6.8
>1 - 2 years	38	18	7.6
>2-3 years	30	14	7.1
>3-5 years	44	21	7.1
> 5 years	48	23	7.0

The table above shows average satisfaction with a server as a function of system age. Satisfaction is rated on a scale of 1 to 10, 1 = completely dissatisfied, 10 = completely satisfied. Users are most satisfied with servers between 1 and 2 years old. The satisfaction rating for the very oldest servers, those 10 years old or older is a surprisingly high 7.1.

With the high overall satisfaction ratings for the vendors represented in this customer base, the median value of those planning to switch primary high availability vendors is a very low 2 (where 1 is *never switch* and 10 is *definitely switch*). Of those surveyed, only 26% have switched high availability server vendors within the last 5 years.

Of the users surveyed, 66% planned on buying additional high availability servers in the next 12-18 months.

Outages & Recovery Times HRG asked users questions about the amount of scheduled and unscheduled downtime their server experienced. HRG also asked what kinds of availability or downtime measurements were kept – specifically with regard to server outages, network outages, end-user outages and the amount of each type of downtime was recorded when available. Further, HRG explored the actual causes of the outages that servers had experienced during the previous year, and how long it took the system to recover.

Downtime by AE

Characteristic	**AE 1**	**AE 2**	**AE 3**	**AE 4**
# of responses	20	29	53	44
Mean (hrs/mo)	8.8	15.	8.9	6.78
Median (hrs/mo)	4.0	5.00	2.0	.83
Downtime (unquantified)	2	21	10	9
Total # of servers in level	25	79	83	69

The preceding table shows the mean and median hours of downtime per month as reported by users. Although users were asked about the exact hours of downtime per month there were some responses that simply noted that there was some downtime – “minimal”, “yes” without any quantification. The number of respondents not quantifying the amount of downtime is shown in the line **Downtime (unquantified).** Unfortunately, we were not able to collect downtime data from all sites. The mean and median are calculated from the number of responses to the question (first line of the table). The last line of the table shows the total number of servers in the Availability Environment. Of the 69 servers in AE 4, 44 gave a numeric answer to the question, 9 indicated they experienced downtime and for 16 sites there was no data available.

The median downtime for all Availability Environments is significantly lower than the mean downtime. Long outages (large numbers), though infrequent, raise the average markedly. Since most sites experience very short outages, this can have a significant impact on user perceptions and clearly causes many users to overrate the Availability Environment of their server.

What Is Measured HRG asked the users we surveyed about the type of outages they measured and their duration. 75% of the users measured outage time for servers; 70% of the users measured network outages and only 45% of the users measured the outage time experienced by the end user.

Outage Data Collected by AE

Type of Measurement	**AE 1**	**AE 2**	**AE 3**	**AE 4**
server outage	77%	72%	78%	75%
network outage	73%	70%	69%	66%
seen by end-user	36%	44%	56%	47%

The table above shows the percentage of sites for each Availability Environment that collect the different types of outage data.

The average duration of server outages is 5.07 hours per month with a median outage of .78 hours per month. Network outages were longer both in average duration and median duration, 5.63 hours per month and 1.25 hours per month respectively. Outages experienced by the end-user had an average duration of 8.22 hours and a median duration of 1.67 hours.

Outages by AE

	AE 1	**AE 2**	**AE 3**	**AE 4**
	Mean /	**Mean /**	**Mean /**	**Mean /**
Type of Outage	**median**	**median**	**median**	**median**
server outage	4.23 /	3.48 /	3.66 /	5.4 /
(hrs/mo)	1.00	.63	.5	.67
network outage	2.01 /	3.71 /	5.69 /	3.68 /
(hrs/mo)	1.64	.83	1.35	.83
end-user outage	NA	NA	6.14 /	5.90 /
(hrs/mo)			2.25	2.08

The preceding table examines the outage time (hours/month) for each class of outage (server, network, or experienced by the end-user) by Availability Environment.

HRG asked users how much unplanned downtime per month they experienced. The following table shows the mean and median number of hours of unplanned downtime per month and the number of responses by Availability Environment.

Unplanned Downtime by AE

Type of data	**AE 1**	**AE 2**	**AE 3**	**AE 4**
# of respondents	16	45	57	51
Mean (hrs/mo)	2.25	1.55	1.93	2.71

Median (hrs/mo)	.75	.50	1.00	.50

You would expect the shortest amount of downtime to be experienced by Availability Environment 4 servers, followed by Availability Environment 3, 2 and 1 servers respectively, however this apparently is not the case.

The data for the following table came from responses to the following questions:

(**Q4**): "Please listen to the following descriptions of high availability environments and tell me which best describes the system you use for your most critical business applications? In the event of a system or component failure the impact on a priority user would be…" The high availability definitions were then read, starting with Availability Environment 4.

(Q5): "Do you measure any of the following: server outage, network outage, outage experienced by the end-user?"

(**Q6**): "What is the average unplanned downtime of your highest availability application server?"

Server Downtime

Variable	**Q4 # of responses**		**Q5 # of responses**		**Q6 # of responses**	
AE4 mean (hrs/mo)	6.78	(44)	5.4	(33)	2.71	(51)
AE4 mediai (hrs/mo)	.83	(44)	.67	(33)	.50	(51)
AE3 mean (hrs/mo)	8.90	(53)	3.66	(40)	1.93	(57)
AE3 median (hrs/mo	2.00	(53)	.50	(40)	1.00	(57)
AE2 mean (hrs/mo)	14.99	(29)	3.48	(30)	1.55	(45)
AE2 median (hrs/mo	5.00	(29)	.63	(30)	.50	(45)
AE1 mean (hrs/mo)	8.77	(20)	4.23	(11)	2.25	(16)
AE1 median (hrs/mo	4.00	(20)	1.00	(11)	.75	(16)

AE4 servers whose mean downtime is 6.78 hours are more than 99% available. An AE4 median downtime of .83 hours turns into a very high 99.9% availability.

What Happened When asked, 61% of the respondents said that they had experienced an outage last year. However, when asked the number of outages experienced last year 69% of the 61% stated they had experienced at least one outage with the median number of outages being two and with the typical outage having a duration of 1.5 hours. The median duration of the worst outage was 3.7 hours while the average duration was 7.94 hours. The difference between the mean and median indicates a few users experiencing perhaps one single very long outage each.

Outage Causes	**# Worst Outages**	**# Typical Outages**
Electrical/power failure (brownouts, lost power, failure of power supplies)	40	25
disk failure	29	13
Hardware failure	26	17
Network or network component (everything but public utility)	8	10
software failure	26	34
application failure	15	16
software version incompatibility	5	0
weather related (acts of God – floods, hurricane)	7	1
lost phone line	3	2
hardware & software	4	7
operator or human error	9	8

The worst outage is on average more than twice the typical outage. Also, causes of worst outages are usually more hardware and environmentally based – 61% for worst outage and only 49% for typical outage. Software errors and human error (operator error) appear to have shorter recovery times than hardware and environmental failures.

HRG asked the respondents what actions they were planning to take to reduce downtime. The responses are shown below.

Steps Planned to Reduce Downtime

Action	**#**	**(%) to take action**
better testing	19	(10%)
RAID	10	(5%)
UPS	26	(13%)
Redundancy	13	(7%)
H/W improvements	27	(14%)
S/W improvements	22	(11%)
no action	30	(15%)
Training	6	(3%)
continuous improvement (review & correct)	13	(7%)
other	30	(15%)

The following table compares the cause of the worst and typical outage with the planned action to reduce downtime.

Outages & Planned Actions to Reduce Downtime

Outage Cause	**# of worst**	**# of typical**	**Number planning**	**Action**

	outage	outage	action	
electrical	40	25	26	UPS
disk failure	29	13	10	RAID
hardware failure	26	17	27	upgrade/ improve H/W
software failure	26	34	22	upgrade/ improve S/W
application failure	15	16	19	better testing
operator/ humai error	9	8	6	training

HRG asked about the level of satisfaction with recovery time. The average satisfaction was 7.3 on a scale of 1 to 10 where 1 is completely dissatisfied and 10 is completely satisfied.

Planned Downtime Upgrades and maintenance on high availability servers were performed during regularly scheduled downtime. The average scheduled downtime per month was 13.94 hours and the median was 4 hours. For many of these servers, being operational 24 X 365 was not a requirement and various types of upgrades and maintenance were performed during the server's scheduled downtime. In fact, 13% of the users surveyed defined high availability as "the system being up when needed – e.g. 6 X 18. The level of satisfaction with the system's scheduled downtime had an average of 9.0 on a scale of 1 to 10 with 1 being completely dissatisfied and 10 being completely satisfied.

Definitions and Concerns The last question on the survey asked what was the respondent's biggest concern with high availability. The table that follows shows the concerns expressed and the percentage of respondents who cited that concern.

HA Concerns

Concern	% concerned
24 X 7 operation	8.8
available to customers & end-users	25.1
fast, accurate recovery	11.9
performance and availability	10.6
data integrity	9.7
disaster recovery	4.8
network operational	4.8

HRG began the interviews by asking the respondents to define what they meant by high availability. The following table shows the responses to that question and the percentage of respondents giving each answer.

Define HA

Define High Availability	% Using Definition
24 X 7	45.5
AE Definitions	7.8
up when needed	14.4
> 99%	10.2

While 24 X 7 operation was the definition of high availability cited by 45% of the respondents only 8.8% stated that 7 X 24 operation was their biggest high availability concern. 22% of the respondents defined high availability in terms of what the end-user experiences ("up when needed" and "the AE spectrum") and consistently, 25% of the respondents stated their biggest high availability concern was being available to their customers and end-users. Although the vendors of high availability servers relate the performance of their product to a 7 X 24 operational requirement, this does not appear to represent the requirements of the majority of high availability server users.

Harvard Research Group, Inc. (HRG)

HRG provides research and consulting services to vendors and users of computer hardware and software products and services.

HRG provides users with a clearer understanding of their options for using information technology.

HRG has real interest in working with computer users to help them make computer technology and architectural decisions most effectively and efficiently, thereby saving them time and money that can be better spent elsewhere in their organizations.

HRG helps its user clients leverage installed technology more effectively while providing a clear view of future IT options.

For more information call

HARVARD RESEARCH GROUP, Inc

Tel. (978)263-3399

Fax (978)263-0033

http//:www.hrgresearch.com