Google Study: Could Those Memory Failures Be Caused By Design Flaws?

By Barbara P. Aichinger, FuturePlus Systems Corporation
JEDEC Memory Server Forum
Shenzhen, China
March 1, 2012

Abstract: The conclusions of the extensive Google study "DRAM Errors in the Wild: A Large-Scale Field Study"[1] revealed that memory failures in the field were far more prevalent than advertised and that no specific conclusion could be reached with regards to the source of the errors. When this landmark study was performed the ability to do real time monitoring of the actual DDR memory was limited, difficult and somewhat costly. Since then the industry has evolved and new technology now exists that can take the Google study to the next level.

Real Time Protocol Compliance violation detection during the live operation of a system has never been achieved in the past due to the inability to monitor the sensitive DDR bus with hardware and software sophisticated enough to do the job. Our dependence on memory subsystems in modern computer architecture makes the validation of DDR subsystems a priority and the ability to quickly find design flaws desirable. Our initial findings using a new tool, the *DDR3 Detective*™[2], show that all the emphasis on the DRAM parts may, for some failures, be pointing the finger in the wrong direction. The sensitive DRAM parts are designed to operate in an environment defined by JEDEC. What happens to these memory parts when the JEDEC specification, which defines how these parts are accessed or how often commands are targeted at them, is outside of the specification? Laboratory and ATE testing stresses the parts with regards to temperature, clock speed and voltage but how will the parts react to actual protocol violations, in the Wild? As the Google study states ***"We found that the incidence of memory errors and the range of error rates across different DIMMs to be much higher than previously reported."*** What Google has found is that laboratory testing and memory system validations used today is sorely inadequate.

What are DDR Protocol Compliance Violations?

JEDEC[3], the industry standard organization that defines the DDR standards, produces timing specifications that govern the protocol of the various DDR standards. A protocol can be thought of as the language that the parts connected to the DDR bus use to talk to each other. Think of it like this: if I am speaking Mandarin to my Chinese customer and I do not say the words correctly, he will misinterpret me and may cancel his order. Thus my inability to speak his language correctly has produced undesirable results.[4] The same is true on the DDR bus. If the protocol is not obeyed, as the chips are designed to expect, they may act in an undesirable fashion.

---

[1] DRAM Errors in the Wild: A Large-Scale Field Study, Schroeder, Pinheiro, Weber; SIGMETRICS/Performance '09 June 15-19 2009, Settle, WA, USA
[2] DDR3 Detective is a trademark of FuturePlus Systems Corporation
[3] www.JEDEC.org
[4] Thank goodness my Chinese customers speak English! ☺

Examples of DDR3 Protocol Compliance Violations

The JEDEC specification tells designers of both memory controllers and DRAM chips what the timing between events can be for correct operation. For the most part these are minimum timings. That is, they do not want events occurring too close together as things are not ready or bus contention can occur.

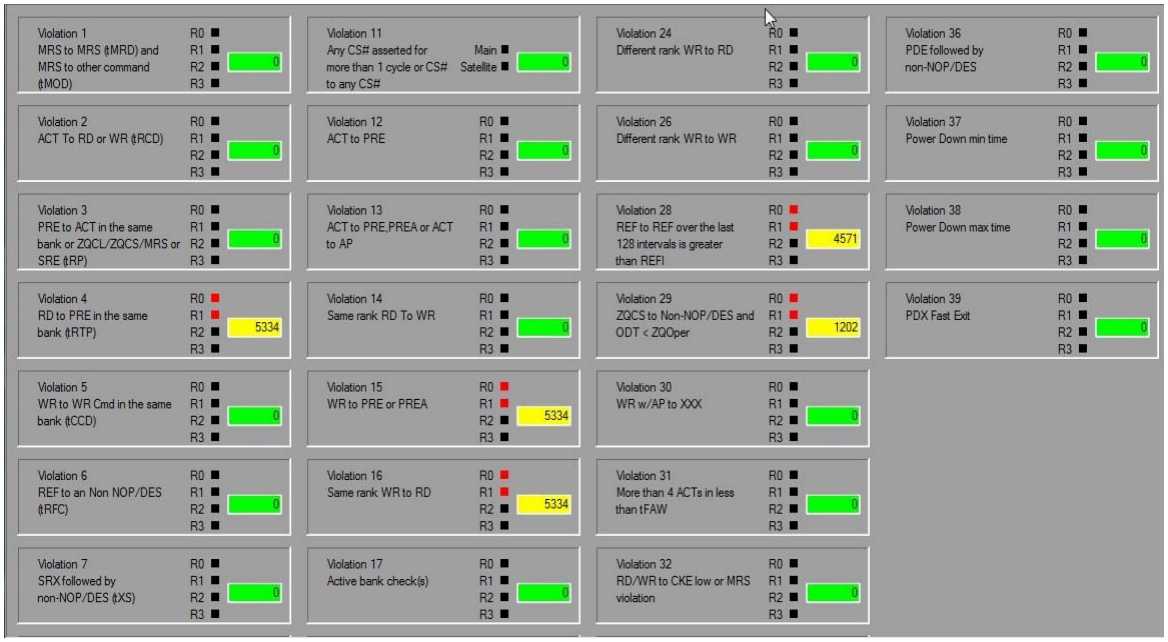


Figure 1: Protocol Compliance Violations found in our system under test. Red indicates failing Rank, yellow indicates number of failures

For example, Section 4.13.3 of the JEDEC standard for DDR3[5] describes a READ operation followed by a PRECHARGE Command. As part of the definition it states "The minimum external Read command to Precharge command spacing to the same bank is equal to AL+ tRTP with tRTP being the Internal Read Command to Precharge Command Delay." So if the system is performing a READ operation it cannot follow that too closely with a PRECHARGE command, which deactivates the open row in a particular bank. In our investigation of a commercially available motherboard we found this to be routinely violated.

[5] JESD79-3E July 2010

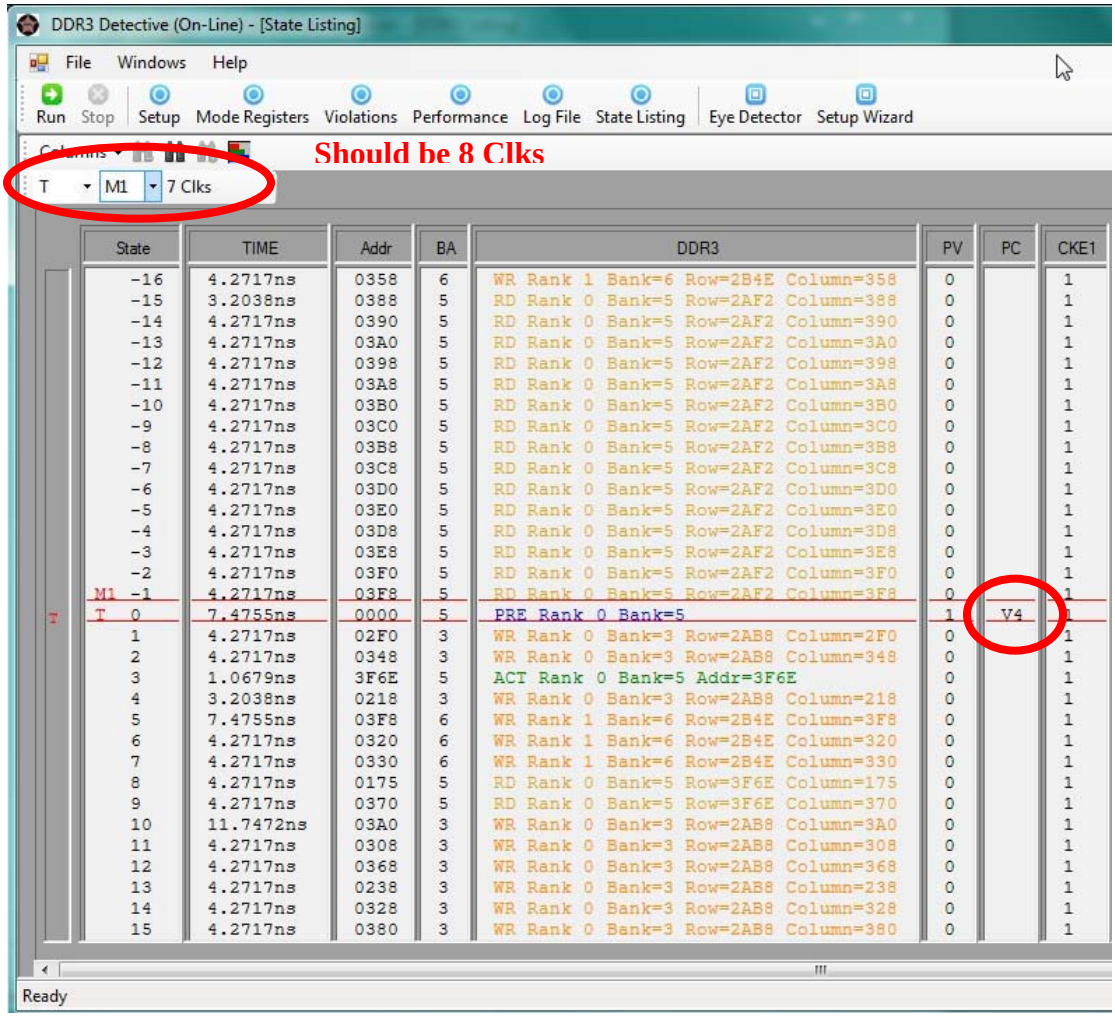


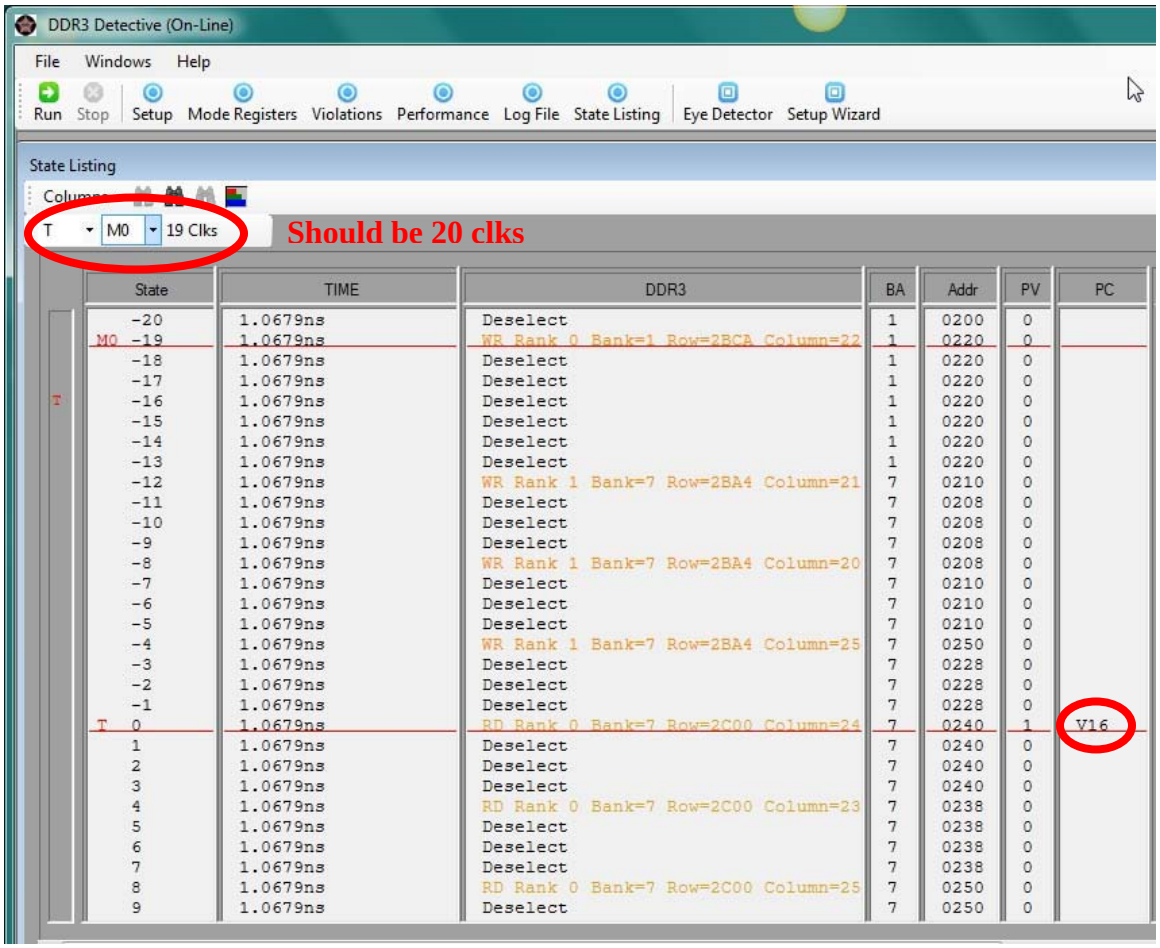
Figure 2: A READ to PRECHARGE Rank 0 Bank 5 separation fails by 1 clock

So what is the possible effect on the DRAM if the system is performing a READ operation too close to Precharging the Rank that the READ operation is targeting? Well if there is a specification saying that you should not do it I don't think you can then blame the DRAM vendor or DIMM vendor if the memory then experiences errors. The engineers performing the Google study did not have access to this type of equipment. However they did conclude: *"We note that, DIMMs within the same platform exhibit similar error behavior, even if they are from different manufacturers."* If different DIMMs from different vendors get the same errors in the same platform perhaps it is not the memory but how the platform treats the memory.

Bus Contention and Catastrophic Errors

The data on the DDR bus is only present on the signal lines for a short period of time. The DDR data bus is shared amongst the different DIMMs in a channel and DRAM parts on a DIMM. It is imperative that once read or write data is on the bus the next read or write data wait until the bus is clear before the new data is put on those same signal lines. This is like a traffic intersection. Don't enter the intersection if there are already cars in that intersection because if the light turns you might experience a collision. A collision of data on the DDR data bus leads to corruption. Some of this corruption is detectable and correctable but some is catastrophic and will result in a system crash or

worse yet undetectable data corruption. The JEDEC specification is detailed in its timing requirements to prevent data collision on the DDR data bus. Even so, we quickly found a WRITE command followed too quickly by a READ command on our system under test.



State	TIME	DDR3	BA	Addr	PV	PC
-20	1.0679ns	Deselect	1	0200	0	
M0 -19	1.0679ns	WR Rank 0 Bank=1 Row=2BCA Column=22	1	0220	0	
-18	1.0679ns	Deselect	1	0220	0	
-17	1.0679ns	Deselect	1	0220	0	
T -16	1.0679ns	Deselect	1	0220	0	
-15	1.0679ns	Deselect	1	0220	0	
-14	1.0679ns	Deselect	1	0220	0	
-13	1.0679ns	Deselect	1	0220	0	
-12	1.0679ns	WR Rank 1 Bank=7 Row=2BA4 Column=21	7	0210	0	
-11	1.0679ns	Deselect	7	0208	0	
-10	1.0679ns	Deselect	7	0208	0	
-9	1.0679ns	Deselect	7	0208	0	
-8	1.0679ns	WR Rank 1 Bank=7 Row=2BA4 Column=20	7	0208	0	
-7	1.0679ns	Deselect	7	0210	0	
-6	1.0679ns	Deselect	7	0210	0	
-5	1.0679ns	Deselect	7	0210	0	
-4	1.0679ns	WR Rank 1 Bank=7 Row=2BA4 Column=25	7	0250	0	
-3	1.0679ns	Deselect	7	0228	0	
-2	1.0679ns	Deselect	7	0228	0	
-1	1.0679ns	Deselect	7	0228	0	
T 0	1.0679ns	RD Rank 0 Bank=7 Row=2C00 Column=24	7	0240	1	V16
1	1.0679ns	Deselect	7	0240	0	
2	1.0679ns	Deselect	7	0240	0	
3	1.0679ns	Deselect	7	0240	0	
4	1.0679ns	RD Rank 0 Bank=7 Row=2C00 Column=23	7	0238	0	
5	1.0679ns	Deselect	7	0238	0	
6	1.0679ns	Deselect	7	0238	0	
7	1.0679ns	Deselect	7	0238	0	
8	1.0679ns	RD Rank 0 Bank=7 Row=2C00 Column=25	7	0250	0	
9	1.0679ns	Deselect	7	0250	0	

Figure 3: Write followed too quickly by a Read to the same RANK

The JEDEC specification for our CAS Latency, CAS Write Latency, and other system parameters dictate that the spacing between these commands to the same RANK should be 20 clock periods. The markers on our *DDR3 Detective*™ Compliance Analyzer show 19.

We did not see any data failures on our MEMTEST but that does not mean that behavior of this type might not lead to failures in the future. Just in case we attached a logic analyzer to the *DDR3 Detective*™ Interposer so that we could see the violation simultaneously with the data. Indeed in the Google Study failures were seen over time and not readily apparent when the systems were first installed.

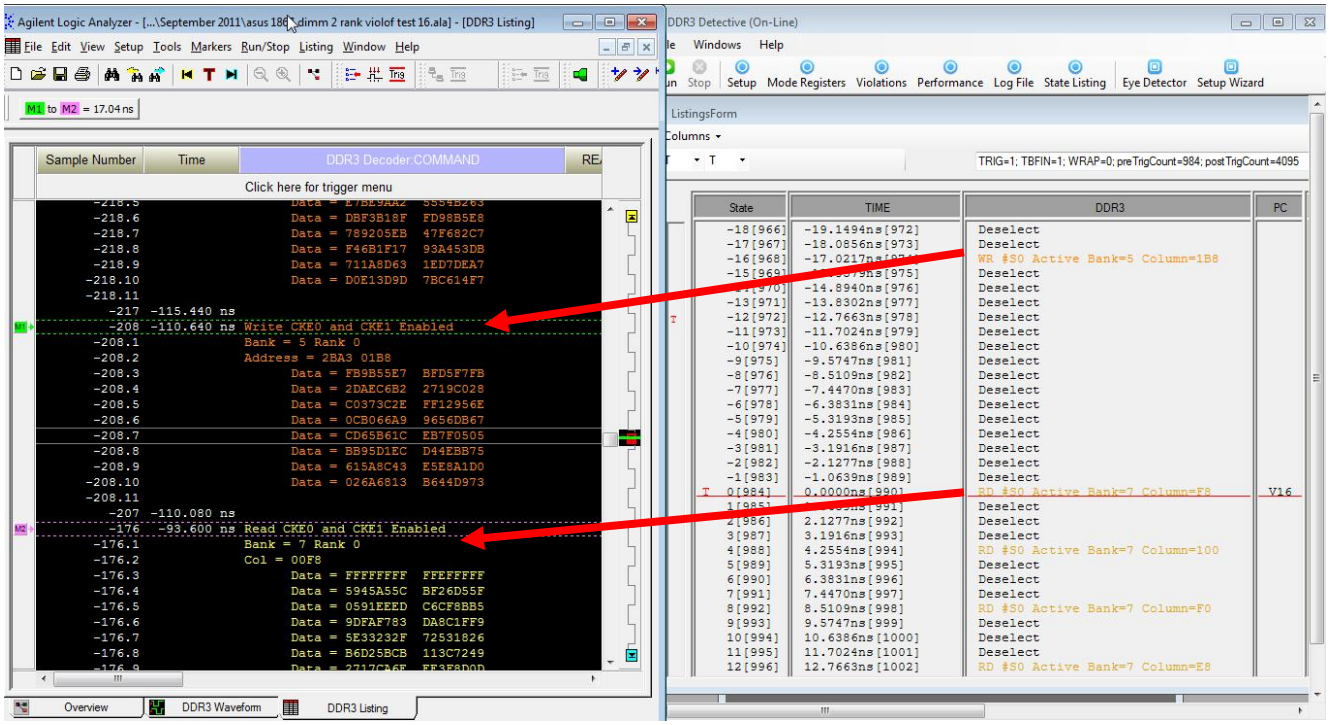
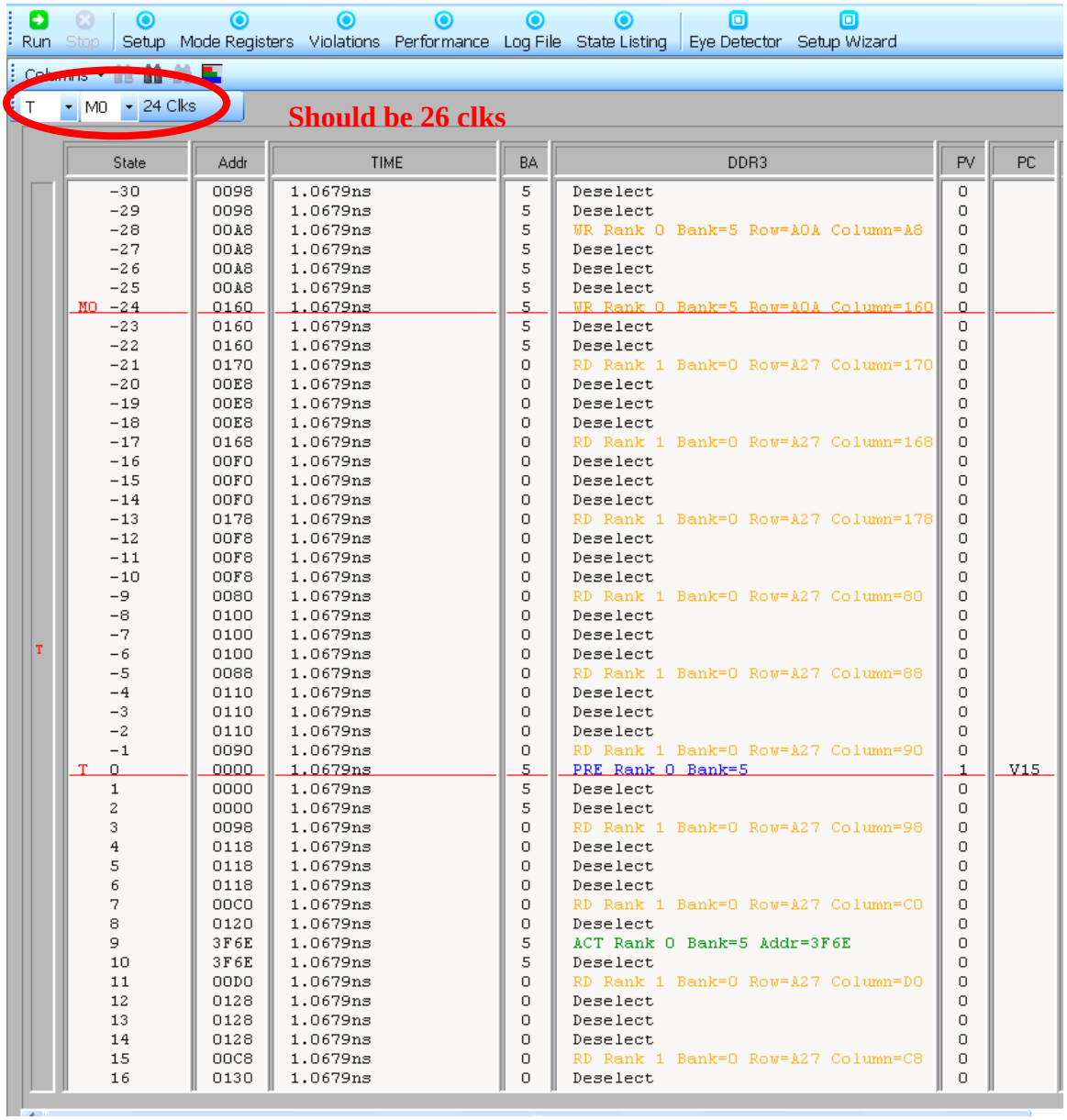


Figure 4: A view of the actual DQ signals for the WRITE command followed too closely by a READ to the same rank (*Logic Analyzer courtesy of Agilent Technologies*)

We also observed WRITE commands followed too closely to PRECHARGE commands to the same bank. Since a PRECHARGE command closes the bank there must be sufficient time for the write data to be written. We can see here a possible corruption of that write data since that the bank is closed too quickly by 2 clocks.



State	Addr	TIME	BA	DDR3	PV	PC
-30	0098	1.0679ns	5	Deselect	0	
-29	0098	1.0679ns	5	Deselect	0	
-28	00A8	1.0679ns	5	WR Rank 0 Bank=5 Row=A0A Column=A8	0	
-27	00A8	1.0679ns	5	Deselect	0	
-26	00A8	1.0679ns	5	Deselect	0	
-25	00A8	1.0679ns	5	Deselect	0	
M0 -24	0160	1.0679ns	5	WR Rank 0 Bank=5 Row=A0A Column=160	0	
-23	0160	1.0679ns	5	Deselect	0	
-22	0160	1.0679ns	5	Deselect	0	
-21	0170	1.0679ns	0	RD Rank 1 Bank=0 Row=A27 Column=170	0	
-20	00E8	1.0679ns	0	Deselect	0	
-19	00E8	1.0679ns	0	Deselect	0	
-18	00E8	1.0679ns	0	Deselect	0	
-17	0168	1.0679ns	0	RD Rank 1 Bank=0 Row=A27 Column=168	0	
-16	00F0	1.0679ns	0	Deselect	0	
-15	00F0	1.0679ns	0	Deselect	0	
-14	00F0	1.0679ns	0	Deselect	0	
-13	0178	1.0679ns	0	RD Rank 1 Bank=0 Row=A27 Column=178	0	
-12	00F8	1.0679ns	0	Deselect	0	
-11	00F8	1.0679ns	0	Deselect	0	
-10	00F8	1.0679ns	0	Deselect	0	
-9	0080	1.0679ns	0	RD Rank 1 Bank=0 Row=A27 Column=80	0	
-8	0100	1.0679ns	0	Deselect	0	
-7	0100	1.0679ns	0	Deselect	0	
T -6	0100	1.0679ns	0	Deselect	0	
-5	0088	1.0679ns	0	RD Rank 1 Bank=0 Row=A27 Column=88	0	
-4	0110	1.0679ns	0	Deselect	0	
-3	0110	1.0679ns	0	Deselect	0	
-2	0110	1.0679ns	0	Deselect	0	
-1	0090	1.0679ns	0	RD Rank 1 Bank=0 Row=A27 Column=90	0	
T 0	0000	1.0679ns	5	PRE Rank 0 Bank=5	1	V15
1	0000	1.0679ns	5	Deselect	0	
2	0000	1.0679ns	5	Deselect	0	
3	0098	1.0679ns	0	RD Rank 1 Bank=0 Row=A27 Column=98	0	
4	0118	1.0679ns	0	Deselect	0	
5	0118	1.0679ns	0	Deselect	0	
6	0118	1.0679ns	0	Deselect	0	
7	00C0	1.0679ns	0	RD Rank 1 Bank=0 Row=A27 Column=C0	0	
8	0120	1.0679ns	0	Deselect	0	
9	3F6E	1.0679ns	5	ACT Rank 0 Bank=5 Addr=3F6E	0	
10	3F6E	1.0679ns	5	Deselect	0	
11	00D0	1.0679ns	0	RD Rank 1 Bank=0 Row=A27 Column=D0	0	
12	0128	1.0679ns	0	Deselect	0	
13	0128	1.0679ns	0	Deselect	0	
14	0128	1.0679ns	0	Deselect	0	
15	00C8	1.0679ns	0	RD Rank 1 Bank=0 Row=A27 Column=C8	0	
16	0130	1.0679ns	0	Deselect	0	

Figure 4: A Write command followed too closely by a Precharge to the same bank

Calibration Commands to the DRAM

The DDR3 JEDEC specification contains the ZQ calibration commands. The purpose of these commands is to perform periodic calibrations to account for voltage and temperature variations. The specification states *"No other activities should be performed on the DRAM channel by the controller for the duration of tZQinit, tZQoper, or tZQCS. The quiet time on the DRAM channel allows accurate calibrations of output driver and on-die termination values."* [6] Even so we quickly found this violation during our investigation. To help us with our analysis of this failure we configured our storage qualification to store only the calibration commands and any protocol violations.

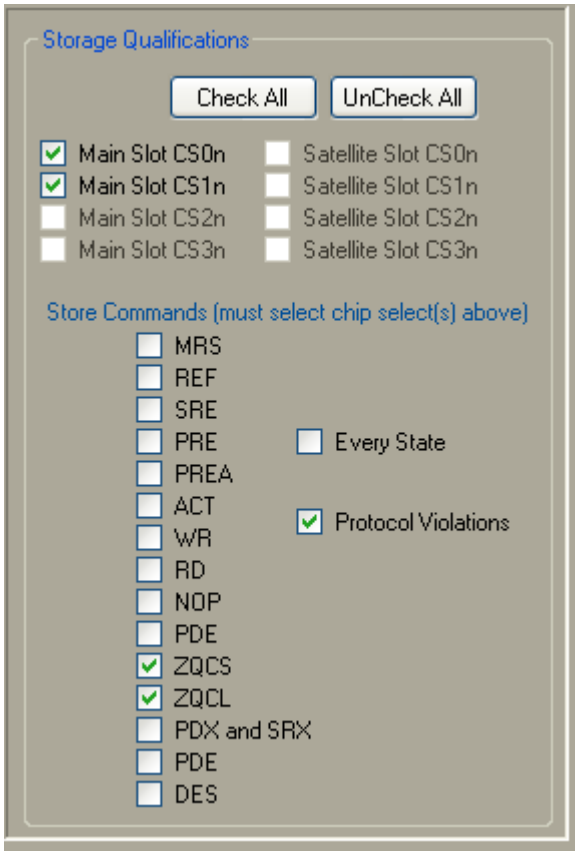
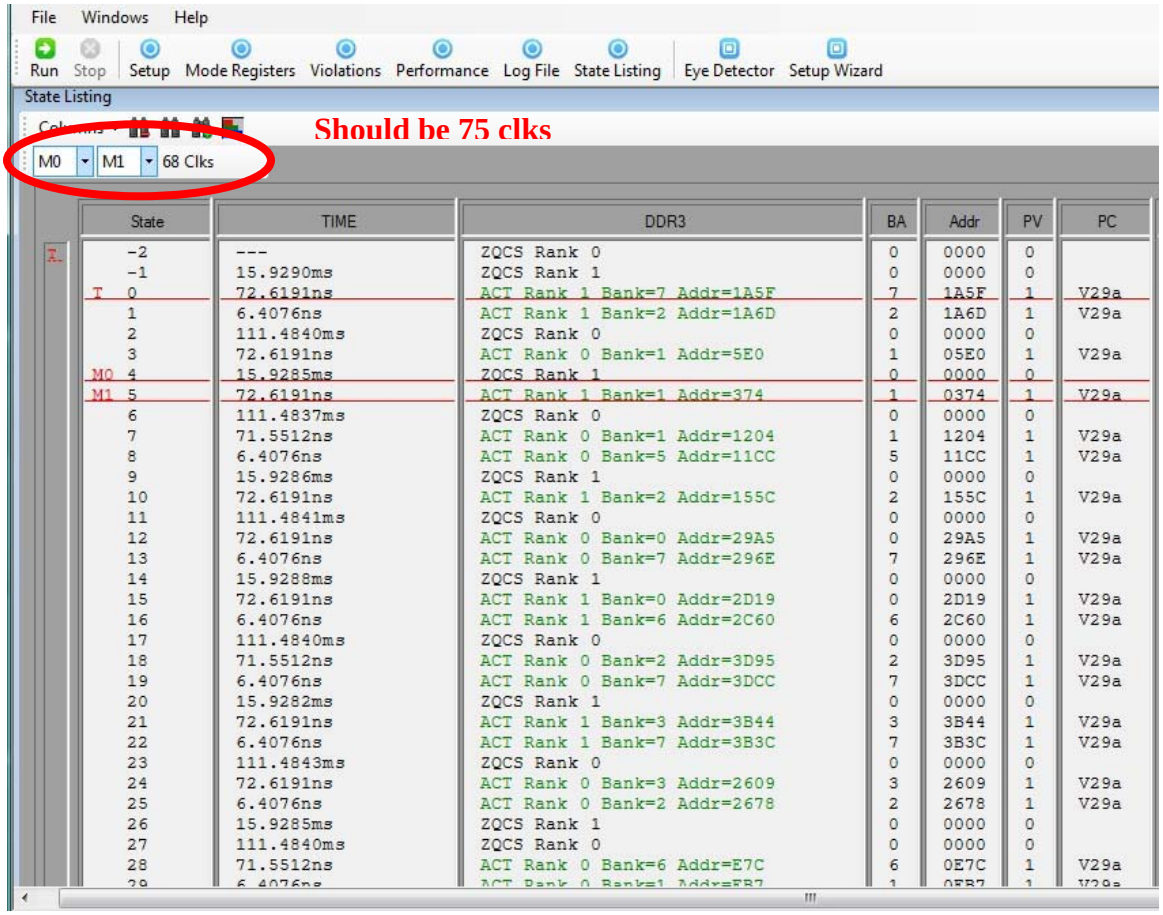


Figure 5: *DDR3 Detective™* Store only protocol violations and Calibration commands

[6] Section 5.5 page 107 JEDEC Standard No.79-3E



	State	TIME	DDR3	BA	Addr	PV	PC
T	-2	---	ZQCS Rank 0	0	0000	0	
	-1	15.9290ms	ZQCS Rank 1	0	0000	0	
T	0	72.6191ns	ACT Rank 1 Bank=7 Addr=1A5F	7	1A5F	1	V29a
	1	6.4076ns	ACT Rank 1 Bank=2 Addr=1A6D	2	1A6D	1	V29a
	2	111.4840ms	ZQCS Rank 0	0	0000	0	
	3	72.6191ns	ACT Rank 0 Bank=1 Addr=5E0	1	05E0	1	V29a
M0	4	15.9285ms	ZQCS Rank 1	0	0000	0	
M1	5	72.6191ns	ACT Rank 1 Bank=1 Addr=374	1	0374	1	V29a
	6	111.4837ms	ZQCS Rank 0	0	0000	0	
	7	71.5512ns	ACT Rank 0 Bank=1 Addr=1204	1	1204	1	V29a
	8	6.4076ns	ACT Rank 0 Bank=5 Addr=11CC	5	11CC	1	V29a
	9	15.9286ms	ZQCS Rank 1	0	0000	0	
	10	72.6191ns	ACT Rank 1 Bank=2 Addr=155C	2	155C	1	V29a
	11	111.4841ms	ZQCS Rank 0	0	0000	0	
	12	72.6191ns	ACT Rank 0 Bank=0 Addr=29A5	0	29A5	1	V29a
	13	6.4076ns	ACT Rank 0 Bank=7 Addr=296E	7	296E	1	V29a
	14	15.9288ms	ZQCS Rank 1	0	0000	0	
	15	72.6191ns	ACT Rank 1 Bank=0 Addr=2D19	0	2D19	1	V29a
	16	6.4076ns	ACT Rank 1 Bank=6 Addr=2C60	6	2C60	1	V29a
	17	111.4840ms	ZQCS Rank 0	0	0000	0	
	18	71.5512ns	ACT Rank 0 Bank=2 Addr=3D95	2	3D95	1	V29a
	19	6.4076ns	ACT Rank 0 Bank=7 Addr=3DCC	7	3DCC	1	V29a
	20	15.9282ms	ZQCS Rank 1	0	0000	0	
	21	72.6191ns	ACT Rank 1 Bank=3 Addr=3B44	3	3B44	1	V29a
	22	6.4076ns	ACT Rank 1 Bank=7 Addr=3B3C	7	3B3C	1	V29a
	23	111.4843ms	ZQCS Rank 0	0	0000	0	
	24	72.6191ns	ACT Rank 0 Bank=3 Addr=2609	3	2609	1	V29a
	25	6.4076ns	ACT Rank 0 Bank=2 Addr=2678	2	2678	1	V29a
	26	15.9285ms	ZQCS Rank 1	0	0000	0	
	27	111.4840ms	ZQCS Rank 0	0	0000	0	
	28	71.5512ns	ACT Rank 0 Bank=6 Addr=E7C	6	0E7C	1	V29a

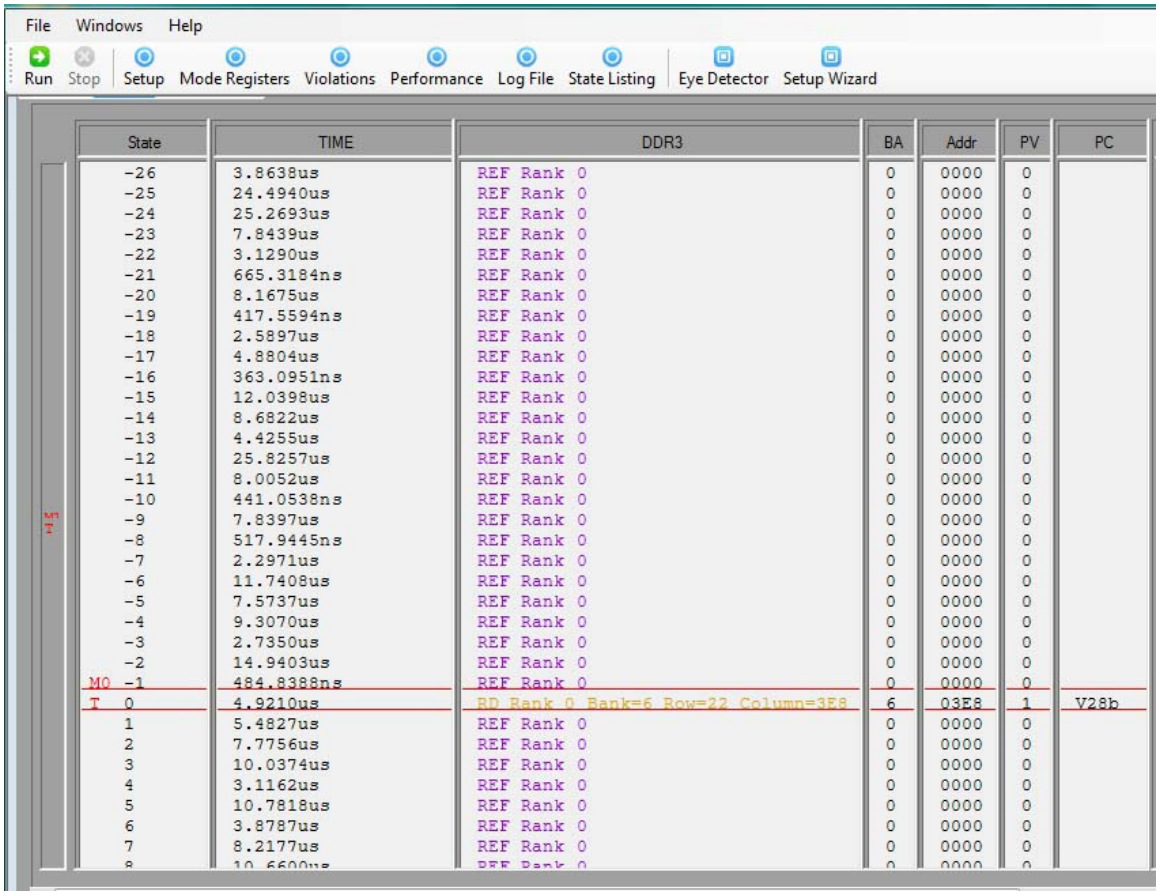
Figure 6: Activate command too soon after a Calibration command.

Observing the time interval we can see how frequently this compliance violation occurs.

Refresh

To maintain the validity of the data the memory controller recharges or refreshes the capacitive cells of the DRAM thousands of times per second. The JEDEC specification contains detailed information on this process as it is critical to maintaining data integrity in the DRAM part. In general, a Refresh command needs to be issued to the DDR3 SRAM every tREFI interval. There is some flexibility allowed for scheduling but at no point in time can more than a total of 8 Refresh commands be allowed to be postponed. Refreshes are also important from a performance perspective. Since dead time is required around a Refresh command one does not want to refresh more than necessary as this wastes memory bandwidth and power consumption. This can be important for server vendors as saving power consumption and improving bandwidth can make the sale.

Below is an analysis of the Refreshes to Rank 0 on our system. When the refresh interval is violated we will see the command that was on the bus when the interval timer for the test expired. Using the store qualification we can see the time interval between Refreshes.



State	TIME	DDR3	BA	Addr	PV	PC
-26	3.8638us	REF Rank 0	0	0000	0	
-25	24.4940us	REF Rank 0	0	0000	0	
-24	25.2693us	REF Rank 0	0	0000	0	
-23	7.8439us	REF Rank 0	0	0000	0	
-22	3.1290us	REF Rank 0	0	0000	0	
-21	665.3184ns	REF Rank 0	0	0000	0	
-20	8.1675us	REF Rank 0	0	0000	0	
-19	417.5594ns	REF Rank 0	0	0000	0	
-18	2.5897us	REF Rank 0	0	0000	0	
-17	4.8804us	REF Rank 0	0	0000	0	
-16	363.0951ns	REF Rank 0	0	0000	0	
-15	12.0398us	REF Rank 0	0	0000	0	
-14	8.6822us	REF Rank 0	0	0000	0	
-13	4.4255us	REF Rank 0	0	0000	0	
-12	25.8257us	REF Rank 0	0	0000	0	
-11	8.0052us	REF Rank 0	0	0000	0	
-10	441.0538ns	REF Rank 0	0	0000	0	
-9	7.8397us	REF Rank 0	0	0000	0	
-8	517.9445ns	REF Rank 0	0	0000	0	
-7	2.2971us	REF Rank 0	0	0000	0	
-6	11.7408us	REF Rank 0	0	0000	0	
-5	7.5737us	REF Rank 0	0	0000	0	
-4	9.3070us	REF Rank 0	0	0000	0	
-3	2.7350us	REF Rank 0	0	0000	0	
-2	14.9403us	REF Rank 0	0	0000	0	
M0 -1	484.8388ns	REF Rank 0	0	0000	0	
T 0	4.9210us	RD Rank 0 Bank=6 Row=22 Column=3E8	6	03E8	1	V28b
1	5.4827us	REF Rank 0	0	0000	0	
2	7.7756us	REF Rank 0	0	0000	0	
3	10.0374us	REF Rank 0	0	0000	0	
4	3.1162us	REF Rank 0	0	0000	0	
5	10.7818us	REF Rank 0	0	0000	0	
6	3.8787us	REF Rank 0	0	0000	0	
7	8.2177us	REF Rank 0	0	0000	0	
8	10.6600us	REF Rank 0	0	0000	0	

Figure 7: A study of tREFI for the system under test

For our system configuration tREFI is approximately 7290 clocks or 7.8us for 1866 operation. We can see in the state listing that there is a wide variation in some of the Refresh intervals. The tool uses a 128 count rolling window of tREFI to calculate the average Refresh Interval.

Measurement Methodology

For this paper we used an Agilent Logic Analyzer, a *DDR3 Detective*™ and a DDR3 memory DIMM interposer.

Figure 8 : When used in conjunction with a logic analyzer the *DDR3 Detective*™ can show the DDR Data bus traffic around the error in addition to the Address/Command/Control signals.

The key to making all this work is that the FuturePlus DIMM interposer sits in the motherboard DIMM socket and unobtrusively taps off the memory bus as it travels to and from the DDR3 memory DIMM that sits atop the interposer. No performance degradation is caused by this interposer. The 'tapped' version of the DDR3 bus signals are then monitored by the *DDR3 Detective*™ logic which runs independent of the logic analyzer.

Figure 9: The *DDR3 Detective*™ unit cabled to the DDR3 DIMM interposer.

A USB link allows the *DDR3 Detective*™ Software to setup the compliance tests and report the results. All the tests can be run without any logic analyzer attached. If the traffic around the violation wants to be observed, such as we have shown earlier, the user can use the internal trace memory or a logic analyzer can be attached. The system under test is unaware of the presence of the *DDR3 Detective*™ and the DIMM Interposer. No special software needs to be run on the instrumented system. Thus the system can run any benchmark software or any application and the memory subsystem can be observed *in the Wild.*

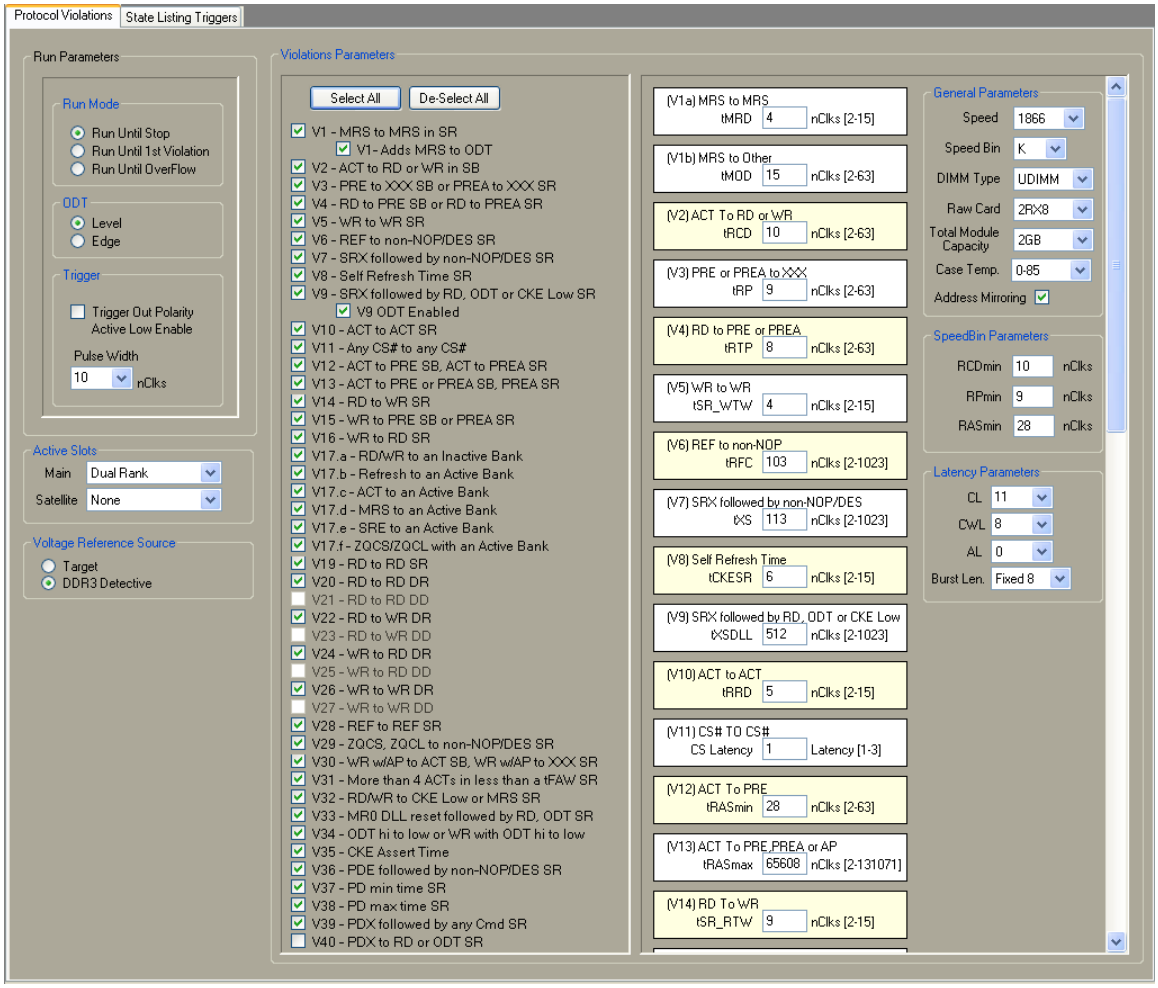


Figure 10 – The *DDR3 Detective*™ Setup Screen auto loads the JEDEC compliance parameters

The Setup Screen allows for the selection or de-selection of over 51 DDR3 Protocol Compliance tests. These tests, across all banks and ranks in the monitored DIMM or SODIMM slot, account for 465 concurrent checks. If systems such as those in the Google study were instrumented with this type of equipment the root cause of some of those failures could have been found.

The Google study was performed from 2006 to 2008 and the results of the study were published in 2009. The industry has moved to DDR3 in the mobile, desktop and server market. Down time and DIMM swapping will continue to prove to be expensive as our society continues its thirst for quick and accurate access to superior information. DDR3 validation is now ready to meet the challenge.

About the Author:
Barbara P. Aichinger holds a Bachelors Degree in Electrical Engineering from the University of Akron, Ohio and Masters Degree in Electrical Engineering from the University of Massachusetts. She is a co-founder of FuturePlus Systems and is currently the Vice President of New Business Development. She is married and has three children.

About FuturePlus Systems:
FuturePlus Systems is an innovator in the Test and Measurements industry and has been in business since 1991. The company has a global customer base and has offices in Bedford, New Hampshire and Colorado Springs, Colorado. The company has representation world wide: http://www.futureplus.com/futureplus-systems-contact-information.html The *DDR3 Detective*™ is the latest in *never been done before* products that FuturePlus prides themselves on. The company is privately owned and can be found on the web at www.FuturePlus.com.