现代系统中的内存错误

摘要

几个最近的出版物表明，内存系统中的硬件故障是常见的。预计这些故障在将来的系统中将会变得更加频繁，这些系统比当前存储器子系统中存在的DRAM和SRAM数量级更多。这些内存子系统在部署在包含数万个节点的高性能计算系统和数据中心时，将需要提供弹性技术来容忍这些故障。因此，了解当前硬件恢复技术的有效性以确定它们是否适合未来的系统至关重要。在本文中，我们介绍了DRAM和SRAM故障和现场错误的研究。我们使用两个领导级高性能计算机系统的数据来分析当前系统中部署的硬件恢复方案的可靠性影响。我们的研究有关于许多目前部署的可靠性技术（如DRAM ECC，DDR地址/命令奇偶校验，以及SRAM ECC和奇偶校验）的功效的几个关键发现。我们还进行方法论研究，发现计数错误而不是故障，研究者和数据中心运营商的常见做法可能导致关于系统可靠性的错误结论。最后，我们使用我们的数据来预测未来大型系统的需求。我们发现，SRAM故障在将来不太可能造成显着更大的可靠性威胁，而DRAM故障将是一个主要关注点，将需要更强大的DRAM恢复能力方案来保持可接受的故障率，类似于当今系统中发现的故障率。

引言

目前的预测是，20世纪70年代初的exascale系统将具有32至100 PB的主要记忆（DRAM），与2012年相比增加了100倍至350倍[8]。在这样的系统中，高速缓冲存储器（SRAM）的量也可能类似地增加。未来的数据中心还将包含比现有数据中心更多的节点。这些系统和数据中心将需要显着增加DRAM和SRAM存储器的可靠性，以便保持与当前系统相当的硬件故障率。这项工作的重点是为系统设计人员和运营商提供可靠系统特性的洞察和指导。特别是，我们的重点是分析现有硬件恢复技术的有效性，对可靠的系统设计的影响，以及它们对未来系统的适用性。为了我们的分析，我们使用了两个领先级生产系统在过去几年收集的数据：Hopper，一个6000节点超级计算机，位于位于奥克兰，卡尔多尼亚的劳伦斯伯克利实验室的NERSC中心; 和位于新墨西哥州洛斯阿拉莫斯州洛斯阿拉莫斯国家实验室（LANL）的一个8,500节点超级计算机的Cielo。这两款超级计算机均基于AMD Opteron TM CPU，并包含DDR3 DRAM。合计，我们分析的数据（可用的全部数据的子集）包含超过3.14亿CPU插槽时数和450亿DRAM设备小时数。该规模使我们有能力收集关于在CPU和DRAM设备的生命周期中观察到的故障的统计代表性数据。虽然我们的数据主要来自超级计算中心，但我们的分析结果适用于硬件可靠性很重要的任何大型数据中心或计算集群。本文增加了关于现有文献的系统责任的几个新颖的见解，并且还突出了分析现场数据的某些方面，这些方面对正确执行至关重要，以获得正确的见解。广泛的贡献包括：

•对Hopper的DRAM和SRAM故障的详细分析。这些数据补充了现有的其他生产系统中的故障研究，包括Jaguar，Cielo和Blue Waters，从而进一步推动了我们对DRAM和SRAM故障的理解[12] [33] [34]。

•对于灾难恢复的影响，对于生产系统来说，这是首次对DRAM的高效性进行分析的研究。•计数错误而不是故障的影响来评估系统的可靠性。计数误差在研究人员和数据中心运营商中是常见的但没有人量化了该方法对获得的可靠性结论的准确性的影响。

•几种基于硬件的弹性技术对系统可靠性的影响，包括SEC-DED ECC，chipkill ECC和DRAM子系统中的命令/地址奇偶校验。我们还研究了SRAM弹性方案，并深入分析设计选择对SRAM可靠性的影响。

•预测SRAM和DRAM故障对未来级别系统的影响。

我们的研究有几个关键的发现，这是系统设计师和操作员感兴趣的。这些发现广泛，覆盖了业界常用的各种硬件恢复技术。我们将这些发现分为三类：好，坏，丑。

好的。我们的研究突出了在理解故障行为和系统可靠性方面取得的进展。这些包括：

•DDR command and address parity，附加到JEDEC的DDR-3和DDR-4规范，对系统可靠性有显着的正面影响，以与ChipKill ECC的未校正错误率相当的速率检测错误。

•现场大多数观察到的SRAM故障都是来自粒子撞击的瞬态故障。这进一步证实了SRAM故障是众所周知的现象和近期未来的工艺技术。

•大多数未校正的SRAM错误都是由于单位故障引起的; 因此，减少SRAM未校正的错误率是工程时间和精力的问题（例如，用ECC替换奇偶校验码），而不是新的研究和新技术。此外，适当的多位故障保护可降低或消除由于高空造成的SRAM故障导致的未校正错误率的预期增加。

坏了  不幸的是，我们的一些研究结果指出需要做更多工作的领域，或者需要更多的理解。这些包括：

•海拔提高了一些DRAM设备的故障率，虽然这种影响因供应商而异。这证明了一些（但不是全部）DRAM器件易受到高能量粒子撞击的瞬态故障的影响。

•与片上SRAM不同，未来系统的外部存储器（例如DRAM）将需要比现有超级计算系统中更强大的恢复技术。

丑陋。最后，我们的一些研究结果显示了常用实践或技术的潜在问题：

•进行实地研究是困难的。例如，我们证明计数错误而不是故障可能导致关于系统可靠性的错误结论。我们检查最近使用错误计数的研究结果，并使用我们的数据来证明使用这种方法可能导致误导或不准确的结论。

•SEC-DED ECC，一种常用的ECC技术，非常适合于现代DRAM子系统，并且可能导致未知的错误（可能导致无声数据损坏）到20FITperDRAM设备，这是许多企业数据中心和高性能计算系统的无法接受的高速率。

本文的其余部分安排如下。

第2节定义了我们在本文中使用的术语。

第3节讨论相关研究，并描述我们的研究和方法学的差异。

第4节解释了Cielo和Hopper的系统和DRAM配置。

第5节描述了我们的实验设置。

第6节介绍了故障的基线数据。

第7节检查计数错误而不是故障的影响。

第8节介绍了我们对现有硬件恢复技术的分析。

第9节从系统设计人员的数据中提取经验教训。

第10节介绍了我们对未来系统可靠性的预测

第11节总结。  
2.术语

在本文中，我们区分故障和错误如下[6]：

•故障是错误的根本原因，例如卡住位或高能量粒子撞击。故障可能发生（导致错误）或休眠（不会导致错误）。

•错误是由活动故障导致的状态的错误部分，例如内存中的错误值。可能会检测到错误，并可能通过更高级别的机制（如奇偶校验码或纠错码（ECC））来纠正错误。他们也可能未经修正，或在最坏的情况下完全未被发现（即沉默）。

硬件故障可进一步分为瞬态，瞬态或硬[7] [10] [11]。区分来自间歇性故障诊断系统的硬故障需要知道准确的存储器访问模式以确定存储器位置是否在每个访问上返回错误的数据。实际上，在我们这样的大规模实地考察中，这是不可能的。因此，我们将间歇性和硬性故障组合在一类永久性故障中。

在本文中，我们研究了各种错误检测和校正机制。其中一些机制包括：

奇偶校验，可以检测到但不能纠正任何单位错误;

单纠错双错误检测错误纠正码（SEC-DED ECC），可纠正任何单位错误并检测任何双位错误; 和chipkill ECC。

我们讨论两个级别的 chipkill ECC：

chipkill-detect ECC，它可以检测但不能纠正单个DRAM芯片中的任何错误;

chipkill-correct，可以纠正单个DRAM芯片中的任何错误。

为了防止多位故障，ECC或奇偶校验的结构有时会使用位交织，这确保物理上相邻的位单元受到不同的ECC或奇偶校验字的保护。

3.相关工作近几年来，已经发表了多篇研究报告，研究生产系统的失败。2006年，施罗德和吉布森研究了LANL超级计算机系统的失败[29]。2007年，李等 发布了三种不同数据集（包括互联网服务提供商的服务器场）的内存错误研究[21]。2009年，Schroeder等 发布了使用Google服务器机队的大规模现场研究[30]。2010年，李等人 发布了一个关于互联网服务器场和其他来源的内存错误的扩展研究[20]。在2012年，Hwang等 发布了关于Google服务器机群的扩展研究，以及两个IBM Blue Gene集群[16]，Sridharan和Liberty介绍了高性能计算系统中DRAM故障的研究[33]和El-Sayed等人。发表了关于DRAM在数据中心环境中的温度影响的研究[14] 。

2013年，Siddiqua等 介绍了客户端和服务器系统中DRAM故障的研究[31]，Sridharan等人 介绍了DRAM和SRAM故障的研究，重点是位置和供应商的影响[3​​4]。最后，在2014年，DiMartinoetal表示了BlueWaters，一个伊利诺伊大学，Urbana-Champaign大学的高性能计算系统[12]。

我们的研究包含比许多先前研究更大的规模和更长的时间间隔，使我们能够提供更多有代表性的故障和错误数据。在具有相似尺度尺度的研究中，我们在第7节中显示的不正确的故障[12] [14] [16] [30]。此外，很少这些以前的研究检查硬件弹性技术如SEC-DED，chipkill和DDR命令/地址奇偶校验的功效。

已经有几十年的DRAM和SRAM的实验室测试（例如[9] [13] [23] [24] [27]），以及Kim等人最近的一项研究。关于DRAM干扰故障[19]。诸如此类的实验室研究允许了解故障模式和根本原因，并补充识别实际发生的故障模式的现场研究。

4.系统配置

我们的研究包括美国两个生产系统的数据：Hopper，位于加利福尼亚州Oakland的超级计算机，高度为43英尺; 和Cielo，位于新墨西哥州洛斯阿拉莫斯的一台超级计算机，海拔约7300英尺。两个系统的相关统计数据摘要见表

1. Hopper包含大约6,000个计算节点。每个节点包含两个12核AMD Opteron TM处理器。每个处理器包括12个32KB的L1数据高速缓存，12个512KB的L2高速缓存和1个12MB的三级缓存。每个节点具有八个4GB DDR3注册DIMM，总共32GB的DRAM。

2. Cielo包含大约8,500个计算节点。每个节点包含两个8核AMD Opteron TM处理器，每个处理器具有8个32KB L1数据高速缓存，8个512KB L2高速缓存和1个12MB L3缓存。每个节点具有八个4GB DDR3注册DIMM，总共32GB的DRAM。

两台机器的节点组织如下。四个节点连接到作为管理模块的时隙。机箱中有8个插槽。三个底盘装在底部到顶部（数字上）inarack. Cielohas 96机架，排列成6排，每个包含16个机架。4.1

DRAM和DIMM配置在Hopper和Cielo中，每个DDR-3 DIMM都包含两个等级18DRAM设备，每个都具有4个数据（DQ）信号（称为x4 DRAM设备）。在每个等级中，16个DRAM器件用于存储数据位，2个用于存储校验位。一条车道是不同级别的一组DRAM设备，共享数据（DQ）信号。同一通道中的DRAM也共享一个选通（DQS）信号，该信号用作数据信号的源同步时钟信号。内存通道有18个通道，每个通道有两个等级（即每个通道一个DIMM）。每个DRAM器件包含可以并行访问的八个内部存储体。逻辑上，每个银行都被组织成行和列。每个行/列地址对标识DRAM设备中的4位字。图1示出了单个存储器通道的图。在物理上，Cielo和Hopper（来自所有供应商）的所有DIMM都是相同的。每个DIMM都是双面的。DRAM器件每侧布置成两行九个器件。Cielo或Hopper的DIMM上没有散热片。两个存储器子系统的主要区别在于Cielo在其存储器子系统上使用chipkill-correct ECC，而Hopper使用chipkill-detect ECC。

5.实验设置

对于我们的分析，我们使用三种不同的数据集 - 来自控制台日志的纠正错误消息，事件日志的未校正错误消息和硬件清单日志。这三个日志提供了将每个错误消息映射到系统中存在于该时间点的特定硬件的能力。

更正的错误日志包含特定时间戳的节点的事件。系统中的每个节点都有一个硬件存储控制器，用于在x86机器检查体系结构（MCA）提供的寄存器中记录更正的错误事件[5]。每个节点的操作系统都配置为每隔几秒轮询MCA寄存器，并将其发现的任何事件记录到节点的控制台日志中。

未校正的错误事件日志类似，并包含系统中记录的未更正错误的数据。这些通常是不经过修正的错误导致的。

控制台和事件日志都包含各种其他信息，包括与每个错误相关联的物理地址和ECC综合信息。使用配置信息进一步解码这些事件以确定与每个错误相关联的物理DRAM位置。对于此分析，我们解码了显示DIMM的位置，以及DRAM库，列，行和芯片。

我们**假设**每个DRAM器件在我们的观察间隔期间经历单个故障。每个DRAM故障的发生时间对应于每个DRAM器件首次观察到的错误消息的时间。然后，我们根据控制台日志中的相关错误为每个故障分配一个特定的类型和模式。我们观察到的故障率表明，在我们的观察窗口内，不到两个DRAM器件会遭受多个故障，因此该方法的误差很小。先前的研究也使用和验证了类似的方法（例如，[34]）。

硬件清单日志是单独的日志，并在其生命周期的不同点提供每台机器中存在的硬件的快照。这些文件提供了对其他系统中经常缺少的硬件配置的独特见解。总共我们分析了400多个个人硬件库存日志，涵盖了Hopper大约三年的时间，在Cielo上两年。这些文件中的每一个包含每个主机硬件的80万到150万行的典型描述，包括配置信息和每个DIMM的信息，如制造商和部件号。

我们将分析限于我们同时具有硬件清单日志以及错误日志的日期子集。对于Hopper，这包括2011年4月至2013年1月的18个月数据。对于Cielo，这包括2011年7月至2012年11月的12个月数据。对于这两个系统，我们排除了头几个月的数据，以避免将我们的数据集开始之前发生的硬故障归因于我们观察的头几个月。另外，在两个系统的情况下，排除了系统处于不一致状态或处于过渡状态的时间段。

为保密目的，我们对所有DIMM供应商信息进行匿名处理。因为这对于CPU供应商来说是不可能的，所以我们以任意单位提供所有CPU数据（即，所有数据都被归一化为图中的数据点之一）。虽然这个说法是绝对的，但它仍然能够观察趋势并比较相对速率。由于每个数据中心和计算系统不同，这通常是从这样的研究中提取的最相关的信息。

5.1方法

Cielo和Hopper在DRAM，L1，L2和L3高速缓存中都包括硬件擦除器。因此，我们可以将永久性故障识别为在洗涤操作中幸存的故障。因此，当设备在多个擦除间隔中产生错误时（即，当错误被包含至少一个擦除操作的时间段分隔）时，我们将故障分类为永久性，并在仅在单个擦除间隔中生成错误时产生错误。例如，Cielo的DRAM擦除间隔是24小时。产生超过24小时的错误的故障被分类为永久性，而仅在24小时内产生错误的故障被归为暂时性故障。

我们观察到的故障率表明，在我们的观察窗口中，不到两个DRAM器件将会遭受多个故障。因此，类似于以前的实地研究，我们做出简化的假设，即每个DRAM设备在我们的观测间隔期间经历单个故障[33]。每个DRAM故障的发生时间对应于来自该DRAM设备的第一次观察到的错误消息的时间。然后，我们会根据控制台日志中该设备的后续错误为每个故障分配一个特定的类型和模式。我们对SRAM故障使用类似的方法（基于故障率）。

6.漏斗故障的基线数据

在本节中，我们介绍了Hopper中DRAM和SRAM故障模式和速率的基线数据。我们已经在以前的工作中发表了关于Cielo的类似信息[34]。

6。1 DRAM故障

图2显示了我们在Hopper测量间隔期间每个供应商的DRAM-小时数。我们的数据涵盖每个供应商大约10亿DRAM小时的运营时间或更多，足以得出统计学意义上的结论。

图3显示了Hopper中随时间的DRAM故障率。与其他系统类似，Hopper在测量间隔期间经历了永久性DRAM故障的下降速率和瞬态故障的恒定速率[33] [34]。

表2显示了Hopper中存在的DRAM故障模式的细分。与以前的研究类似，我们确定了几种独特的DRAM故障模式：

单位，其中所有错误映射到单个位;

单字，其中所有错误映射到单个字;

单列，其中所有错误映射到单个列;

单行，其中所有错误映射到单个行;

单一银行，其中所有错误都映射到单一银行;

多银行，其中错误映射到多个银行;

多等级，其中错误映射到同一车道中的多个DRAM。

与其他DDR-2和DDR-3系统类似，大多数DRAM故障都是单位故障，但是不平凡的故障是多位故障，包括行，列，存储和芯片故障。

图4显示了由DRAM供应商细分的每个DRAM器件的FIT中Hopper的瞬态和永久故障率。Hopper A，B和C在Cielo [34]中与A，B和C中的A，B和C相同。该图显示了每个供应商FIT率的显着变化，与Cielo的数据一致[34]。这意味着DRAM供应商的选择是系统可靠性的重要考虑因素。

6.1。高海拔的影响

众所周知，由于宇宙辐射的高能中子，高度将对现代SRAM器件产生影响[7]，而DRAM中的粒子撞击是过去的一个问题[23]。然而，高能中子是否是现代DRAM器件中的重大故障来源是不确定的。我们以前在Cielo发表了关于DRAM故障的数据[34]。我们将这些数据与Hopper数据结合使用，通过比较Cielo和Hopper中类似DRAM设备的故障率来检查高度对DRAM故障的影响。

Hopper和Cielo存储器子系统非常相似：系统使用来自同一三个DRAM供应商的内存，共享相同的DDR-3内存技术，并使用相同的内存控制器。两个系统之间的主要区别在于它们的高度：中子通量，相对于纽约市，Cielo经验为5.53，Hopper经历的相对中子通量为0.91 [1]。因此，通过比较Cielo和Hopper的故障率，我们可以确定高度对DRAM故障率有什么影响（如果有的话）。

图5绘制了Cielo瞬态故障率与Hopper瞬态故障率的关系，采用Cielo数据[34]。该图显示，供应商A体验到Cielo的虚拟变化，而B经历了适度更高的瞬态故障率，供应商C经历了几乎相同的瞬态故障率。

表3分解了这些数据，并显示了DRAM厂商的每个系统的单位，单列和单行瞬态故障率。该表显示，供应商A和B的单位瞬态故障率在Cielo中显着高于Hopper，供应商A的单列和单组瞬态故障率也是如此。所有其他瞬态故障模式的速率在两个系统之间的1 FIT内。由于两个系统之间的相似之处，Cielo中更高的单位，单列和单组瞬态故障率的最可能原因是高能中子的粒子撞击。

我们从这些数据的主要观察是，粒子诱导的瞬态故障在DRAM器件中仍然是显着的影响，尽管敏感性因供应商而异，并且在现代DRAM中存在明显的其他故障原因（包括瞬态和永久性）。此外，我们的研究结果表明，虽然高度确实对系统的可靠性有影响，内存供应商的多余选择可以减少这种影响。

6.1.2多位DRAM故障

图6显示了具有位翻转的多个相邻DRAM行的Hopper的单组故障的例子。该图显示，来自该故障的错误限于三个逻辑相邻行中的多个列。我们的数据集中的多个故障与这种一般模式相匹配，尽管发生率相对较低。所有错误均由ECC纠正。

这种故障的错误模式似乎与称为DRAM干扰故障或“行锤”故障的故障模式相似[19] [26]。当在相对较短的时间窗口中反复打开“侵略者行”时，此故障模式会导致“受害者行”的破坏。我们没有对这部分进行根本原因分析; 因此，我们不能肯定地说，故障是一个干扰故障。然而，很明显，具有相似特征的故障模式确实在实践中发生，必须予以说明。

我们从本节数据的主要观察是，在系统架构，设计和甚至部署之后，可能会出现新的和意想不到的故障模式：行锤故障仅在Hopper设计后才被行业识别，并且可能在服务开始[26]。先前的工作已经考虑到对特定故障模式的定制检测（例如[35]），但是数据显示，我们需要能够处理更多种类的故障模式，并且鲁棒的检测至关重要。

6.2 SRAM故障

在本节中，我们将讨论Hopper中AMD Opteron TM处理器中SRAM的故障率。

6.2。1 SRAM Fault Rate图7显示了Hopper中每个插座的SRAM故障，由硬件结构分解。该图清楚地表明，SRAM故障率由处理器中两个最大结构的L2和L3数据高速缓存中的故障所主导。然而，很明显，许多结构中出现故障，包括较小的结构，如标签和TLB阵列。

该图的一个关键发现是，即使在片上结构较小的情况下，故障也会发生，并且这些结构中的检测和校正对于确保在大型系统上运行的高性能计算工作负载的正确性至关重要。这个结果对于系统设计人员来说很重要，因为考虑到不同的设备，基于他们的系统。设备应该在所有SRAM上提供可靠的覆盖和顺序逻辑的重要部分，或提供替代的保护手段，如冗余执行[36]。

6.2.2与加速测试的对比

加速粒子测试通常用于确定SRAM器件对能量粒子的单事件扰动的灵敏度。要综合考虑，加速测试必须包括SRAM单元在现场暴露的所有颗粒（例如，中子，α粒子），其能量近似于现实条件。因此，重要的是将加速测试数据与现场数据相关联，以确保条件实际上类似。

在Hopper使用的SRAM上的加速测试使用各种粒子源进行，包括LANL上的高能中子束（LANSCEICE House）[22]和α粒子源。测试是“静态”测试，而不是运行测试。静态测试用已知值初始化SRAM单元，将其暴露给光束，然后将结果与初始状态进行比较以查找错误。

L2和L3高速缓存使用硬件擦除器，因此我们期望字段错误日志捕获在这些结构中发生的大多数位翻转。图8将Hopper L2和L3数据和标签阵列中的SRAM故障的每比特率与SRAM单元上的加速测试活动相结合的结果进行比较。该图显示，加速测试预测故障率低于L2标签阵列以外的所有结构中的故障率。L2标签中的故障率大约等于加速测试的速率。总体来说，该图显示了在现场测量的速率与从静态SRAM测试测量的速率之间的良好相关性。我们从这种相关性的结论是，现场的SRAM故障的大部分是由已知的粒子引起的。

虽然预期的结果，确认现场数据的期望对于确保部件按照规定运行是重要的，以确定潜在的新的或意外的故障模式，这些故障模式可能没有在预生产硅胶中进行测试，并确保加速测试反映现实。

7.测量系统健康方面的谬误

上一节中提供的所有数据和分析都涉及故障率，而不是错误率。一些以前的研究报告了系统错误的突破[12] [14] [16] [30]。错误率在很大程度上取决于系统的软件配置及其工作负载的访问模式以及系统的运行状况，这使得错误率成为硬件可靠性的一个完美的衡量标准。在本节中，我们显示测量误差率可能导致关于硬件可靠性的错误结论。

7.1错误记录架构

在大多数x86 CPU中，DRAM错误记录在北桥区块的寄存器组[4] [31]中。每个寄存器组一次可以记录一个错误; x86架构规定硬件会丢弃随后的纠正错误，直到操作系统读取和清除存储区[5]。操作系统通常通过每几秒执行一次轮询程序读取寄存器组[2]。因此，在每纳秒发出多个存储器访问的处理器上，在寄存器组的连续读取之间可能会丢弃数百万个错误。

上述错误记录体系结构意味着控制台日志仅代表在系统上发生的所有已更正错误的示例。此外，记录的校正错误的数量高度依赖于由操作系统设置的轮询频率。例如，如果操作系统使用5秒轮询频率，则每5秒间隔只能报告每个节点一个错误。

另一方面，x86处理器中的未校正错误通过单独的机器检查异常机制[5]收集。这些异常被单独传递到操作系统，因此未校正的错误计数不会遭受该采样问题。

7.2计数错误与计数错误

使用我们的数据，我们提供了两个具体的例子，说明计数错误如何导致关于系统可靠性的错误结论。我们的第一个例子是Di Martino等人的研究 在蓝色水系统[12]。第二个例子是Schroeder等人的研究 关于Google数据中心的内存错误[30]。

示例1.首先，我们研究Di Martino等人 声称在Blue Waters的DRAM上使用的chipkill ECC校正了所有记录错误的99.997％。因为校正错误是在未校正错误的情况下进行采样的，因此Blue Waters上的chipkill代码几乎肯定会纠正系统遇到的总错误的大部分。然而，更重要的是，这种分析可能导致关于ECC功效的错误结论。例如，使用作者的方法，我们发现Hopper的chipkill-detect ECC校正了所有记录错误的99.991％，而Cielo的chipkill-correct ECC校正了所有记录错误的99.806％。这意味着Hopper的ECC性能优于Cielo的ECC。然而，Cielo的ECC实际上具有比Hopper低3倍的未校正错误率，导致了相反的结论：Cielo上的chipkill-correct代码的性能比Hopper上的chipkill-detect代码好得多。

示例2.我们的第二个例子来自Schroeder等人的一篇论文，其中作者引用了25,000-75,000 FIT / Mbit的记忆错误率，并声称DRAM比以前认为的可靠性要低得多[30] 。作者测量的速率是记录错误率，而不是故障率[28]。使用这种方法，我们发现Hopper的内存错误率略高于Cielo的4倍，再次给人的印象是，Hopper的DRAM不如Cielo的可靠性（见图9）。然而，如第6节所述，与Cielo的40 FIT /设备相比，Hopper的故障率为25 FIT /设备[34]，表明Hopper的DRAM实际上比Cielo更可靠。

7.3计数故障的重要性

本节的关键观察是为了获得系统可靠性的准确图像，必须分析错误日志以识别单个故障，而不是将每个错误事件分开处理。错误事件计数：（a）更多地表示操作系统的查询频率，而不是硬件健康状况; 和（b）过份强调永久性故障的影响，这可能导致数千或数百万的误差，同时报告瞬态故障的影响，这通常只会导致一些错误，但对系统可靠性同样有害。

8.硬件弹性方案分析

在本节中，我们使用我们的数据来检查DRAM和SRAM中的各种硬件恢复方案。我们的目标是了解当前硬件方案的有效性。这个数据系统设计人员应该考虑一系列的弹性技术。

8.1比较DRAM ECC方案

业界采用了许多DRAM ECC方案。最常见的方案是SEC-DED ECC，chipkill检测ECC​​和chipkill-correct ECC [18]。SEC-DED ECC纠正单位错误并检测双位错误。

Chipkill检测ECC​​可以从单个DRAM设备检测到任何错误，而chipkill-correct ECC可以纠正单个设备的任何错误。先前的研究表明，chipkill-correct将未校正的错误率相对于SEC-DED ECC减少了42倍[33]。然而，以前的研究没有量化SEC-DED和chipkill之间未检测到的差异的差异。

Cielo和Hopper都会针对每个更正的错误记录错误位。通过分析每个错误，我们可以确定给定的错误是否被SEC-DED ECC检测不到。为了本研究的目的，**我们假设在单个ECC字中大于2位的错误被SEC-DED检测不到（实际的SEC-DED ECC通常可以检测到这些错误的一部分）。**我们称之为在ECC字中产生大于2位错误的故障，即SECDED故障。如果在任何ECC字中影响超过两位，并且写入该位置的数据与故障产生的值不匹配，则通过SECDED检测到故障。例如，将1写入具有卡住1故障的位置不会导致错误。

并非所有多位故障都不能被SECDED故障检测到。例如，许多单列故障仅影响每个DRAM行的单个位[33]，并且显示为具有可通过SEC-DED ECC校正的单位错误的一系列ECC字。

图10显示了由供应商细分的Cielo不可检测的SECDED故障率。这些缺陷的发生率对供应商的依赖性很强。供应商A每个DRAM设备的SECDED故障率为21.7 FIT。另一方面，供应商B和C分别具有1.8和0.2 FIT /器件的低得多的速率。Cielo节点具有288个DRAM器件，因此分别为供应商A，B和C分别提供6048,518和57.6个FIT节点。这意味着每隔0.8天，每隔9.5天发生一次未检测到的错误，并且每隔85天在Cielo的大小上进行。图10还显示，在产生较大（例如3位）误差之前，Cielo上30％的不可检测的SECDED故障产生2位错误（可由SEC-DED检测），而剩余的70％不。然而，我们强调，这种检测不是保证 - 不同的工作负载会将不同的数据写入存储器，从而潜在地表现出不同的多位错误模式。

我们从这些数据的主要结论是SEC-DED ECC不太适合现代DRAM子系统。未检测到错误的速率太高，无法证明其在包含数千个节点的大型系统中的使用，其中结果的忠实度至关重要。即使是最可靠的供应商（供应商C），在考虑并行运行的多个节点时，也可以通过SECDED故障检测到高速率。Hopper和Cielo分别使用chipkill-detect ECC和chipkill-correct ECC，因此与使用SEC-DED ECC相比，显示出低得多的未检测到的错误率。

8.2 DDR命令和地址奇偶校验

DDR3（现在的DDR4）存储器的一个关键特性是可以向命令和地址总线添加奇偶校验逻辑。该总线上的电线从存储控制器共享到每个DIMM的寄存器。因此，DIMM上的所有DRAM设备都会看到这些电线上的错误。**虽然DDR2内存系统上的命令和地址奇偶校验是可选的，但我们并没有预先研究这种奇偶校验机制的潜在价值。**

on-DIMM寄存器计算接收到的地址和命令引脚的奇偶校验，并将其与接收到的奇偶校验信号进行比较。在不匹配时，寄存器向存储器控制器发送奇偶校验错误信号。该标准无法在检测到的奇偶校验错误时重试，但要求DIMM内存寄存器不允许任何错误的事务将数据写入存储器，从而防止数据损坏。

Cielo中的DDR3子系统包括命令和地址奇偶校验。图11显示了相对于检测到的未校正的数据ECC错误的速率，检测到的命令/地址奇偶校验错误的速率。该图显示了命令/地址奇偶校验错误率是未校正ECC错误率的72％。该数据的结论是，命令/地址奇偶校验是DDR标准的有价值的补充。此外，增加DDR内存通道速度可能会导致信号相关错误的增加。因此，我们预计地址奇偶校验与ECC错误的比率随着DDR频率的增加而增加。

8。3 SRAM错误保护

在本节中，我们从Hopper和Cielo中检查SRAM中未校正的错误。

图12显示了任意单位的Cielo上SRAM未校正错误的速率。该图分别描述了来自ECC保护结构的奇偶校验保护结构和未校正错误的未校正错误，并包括核心，所有高速缓存以及各种非核心阵列和FIFO中的结构。ECC保护的结构包括L1，L2和L3高速缓存，其包括处理器中的大部分管芯区域。该图显示，Cielo中大多数未纠正的错误来自平价保护结构，尽管其结构与ECC保护结构相似。奇偶校验可以检测到但不能纠正单位故障，而ECC将纠正单位故障。因此，Cielo中SRAM未校正错误的主要原因是单位而不是多位故障的结果。

我们在本节中的主要结论是，减少SRAM未校正错误率的最佳方法是通过处理器中的其他结构来扩展单位校正（例如，SEC-DED ECC）。虽然这是由于性能，功率和面积方面的问题而进行的，但这种解决方案并不需要大量新的研究或新技术。一旦解决了单位故障，剩余的多位故障可能更为挑战，特别是在高比例过程技术中，多位故障的速率和扩展可能会大大增加[17]。此外，诸如超低电压操作的新技术可能会改变硅破坏特性[37]。然而，当前和将来的工艺技术将从增加片上ECC的使用中获益匪浅。

8.4 SRAM错误分析

在本节中，我们深入了解观察到的SRAM未校正错误的细节，以确定缓解的根本原因和潜在途径。

图13显示了来自Hopper中几个代表性的SRAM结构的未校正误差（以单位为单位）的速率。在Hopper中使用的处理器中，L1数据缓存标签（L1DTag）受到奇偶校验保护，所有错误都不可修正。另一方面，L2缓存标签（L2Tag）在很大程度上被ECC保护。但是，每个L2标签条目中的一个位都是由奇偶校验保护的。由于L2高速缓存大大超过L1，在L2标签中有大约一半的奇偶校验位，因为整个L1标签中有位。这反映在来自L2标签的观察到的未校正错误率中，大约是来自L1标签1的未校正误差的一半。我们从这些数据的观察结果是，看起来很小的微架构决策（例如，从ECC保护中排除一点的决定）可能会对整个系统故障率产生很大的影响。因此，微架构的详细建模和分析（例如，AVF分析[25]）对于确保弹性系统设计至关重要。

事实上现场研究表明，由于大气中子的影响，SRAM故障率具有很强的高度依赖性[34]。对于这种趋势，Cielo比Hopper经历了更多的更正的SRAM故障。虽然第一个原则可以预测高度上SRAM未校正误差的类似增加，但是以前的研究还没有量化是否在实践中是正确的。

图14显示了在L1指令标签，L2标签和L3数据阵列中，Cielo相对于Hopper的每位SRAM未校正错误率。该图显示，尽管处于较高的高度，但L2和L3结构中的Cielo错误率并未显着高于Hopper的误码率。我们将其归因于这些结构中的错误保护。这些结构具有ECC和积极的位交织;

因此，需要大于2位的击穿才能导致未校正的错误。因此，在这些结构中，高能中子的数据点与多位误差的比较小于其他来源的误差率，如Alpha粒子，以及各种硬故障机制。

因此，本节中的数据表明，适当的错误保护机制可以成功地抵消由于高度导致的原始故障的增加。我们得出结论，当位于较高的高度时，适当设计的系统不需要较不可靠。

9.可靠的系统设计课程

从我们的研究中收集到有关可靠的系统设计的几个经验教训。我们认为这项研究既确认并反驳了许多广泛的假设，也为系统设计师和研究人员提供了有价值的新见解。

•故障是不可预知的。某些故障可能比预期更频繁（例如，DDR地址奇偶校验），并且某些故障模式在设计时可能不知道（例如，DRAM干扰故障）。因此，提供强大的错误检测器是关键

•细节问题。例如，将缓存简单地描述为“ECC保护”不能传达足够的信息，因为从该保护中排除甚至一个位可能会对系统可靠性产生很大的负面影响。需要仔细分析和建模（如AVF分析）来预测任何设备的预期故障率。

•诊断能力至关重要。这包括确保硬件日志具有足够的诊断信息以及访问适当的软件工具。例如，确切知道哪些位是错误的，这对于了解故障模式和确定出错的情况至关重要。在硬件上增加这种诊断水平可能需要比将ECC添加到结构中的时间和精力更大的投入。

•分析是棘手的。从本质上讲，我们的经验是，这种类型的实地考察难以正确执行。这些研究需要了解硬件和软件行为的细节（其中一些可能是未记录的），挖掘非常大的数据集，以将信号与噪声分开，并仔细解释结果。其次，我们的演示不再是这些步骤可能导致关于系统可靠性的错误结论。

•比例改变一切。在大型系统中，即使非常小的结构也需要保护，因为现代超级计算机或数据中心中的组件数量庞大。忽略这种类型的保护的组件可能会以非平凡的速率破坏数据。

10.对未来系统的预测

在本节中，我们将研究DRAM和SRAM故障对潜在的超级计算机的影响。我们的目标是了解现有的可靠性机制是否可以应对未来大型系统和数据中心可能会增加系统容量和故障率所面临的挑战。为了实现这一点，我们将观察到的DRAM和SRAM错误率扩展到可能的exascale配置和系统大小。我们还模拟了不同工艺技术（如FinFET晶体管）的影响[15]。

本节中的所有分析均指检测到的错误。我们的预测是针对当前存在的技术，不考虑新技术的影响，如堆叠式DRAM，超低电压CMOS或NVRAM。

10.1 DRAM

Exascale超级计算机预计在主存储器的32PB和128PB之间。由于芯片堆叠器件的容量限制[32]，大部分存储器可能会在片外存储器中提供。这些片外器件的接口可能类似于当前的DDR存储器接口。因此，了解这些片外子系统的可靠性要求至关重要。

先前的工作表明，尽管功能尺寸减小和密度增加，DRAM供应商在技术世代中保持大约每个设备的故障率[9]。因此，我们预计超级计算机中的每个设备故障率将类似于当今DRAM设备中观察到的故障率。我们的目标是确定当前的纠错码（ECC）是否足以支持超级计算机。

因此，我们假设exasc系统中的每个记忆体通道将使用Cielo上使用的相同的单芯片代码。因此，我们预计，超级计算机中的每个设备未校正的错误率将与Cielo中的每个设备错误率相同。因为我们不知道在Exascale时间尺度上大量生产的确切的DRAM器件容量，所以我们将设备容量从8Gbit到32Gbit的设备进行扫描。更大的DRAM设备可以通过更少的总体设备提供指定的系统内存容量。

图15显示了该分析的结果。该图描绘了在Cielo上相对于DRAM未更正错误率的exascale系统的系统级DRAM未校正错误率。该图在每个点绘制了每个设备的两个错误率，每个都取自不同系统上的现场数据。该图显示，超高压系统的未校正误差率范围为高端3.6倍Cielos未校正错误率至69.9倍Cielos未校正误差率。在系统层面上，内存容量高端DRAM未修正错误率的增长与从SEC-DED ECC升级到芯片级ECC [33]时实现的DRAM未校正错误的40倍减少相当。如果我们假设SEC-DED ECC为今天的内存子系统提供了不足的可靠性，我们从这些结果得出的结论是，高容量exascale系统可能需要比chipkill更强的ECC。

10.2 SRAM

我们现在将注意力转移到未来系统上的SRAM故障。Cielo中的一个插座在L2和L3高速缓存数据阵列中包含18MB的SRAM。Exascale级处理器预计将显着增加每个插座的处理能力，因此每个节点将显示更多的SRAM。考虑到CPU和GPU处理器的技术趋势和报告的结构尺寸[3]，我们预计，一个exascale插座将包含超过150MB的SRAM，或者每个插座的SRAM在当前的电脑上增加8-10倍。由于切换到加速处理单元（APU）中与CPU内核集成的通用图形处理单元（GPG-PU）和/或GPGPU，SRAM相对于今天系统的增加小于DRAM的增加），这两者都比传统CPU内核更少依赖于大型SRAM缓存。假设SRAM故障率在未来几年保持不变，每插槽故障率将随SRAM容量线性增加。因此，exascale处理器将看到当前处理器遇到的SRAM故障数量的8-10倍，而未修正的SRAM错误的速率可能会达到8-10倍。这意味着系统级的未校正错误率来自Cielo上SRAM未校正错误率的50-100倍的SRAM错误，具体取决于系统大小。这在图16的第一组条中示出，对于低节点计数系统（50k个节点）标记为小，对于高节点计数系统（100k个节点），标记为小。近年来，每位SRAM瞬态故障率呈下降趋势[17]。如果这种趋势继续下去，每个插座的SRAM未校正错误率将低于我们的预期。例如，根据Ibe等人 45nm至22nm技术的每位SER降低了62％。如果我们看到当前的CMOS技术和exascale技术相应的减少，则exascale系统SRAM未校正的错误率降低到Cielo的未校正错误率的19-39倍，如图16的第二组条所示。最后，如前所述，大多数未校正的SRAM错误是由于单位故障。如果这些错误在exascale处理器中被消除（例如，通过用ECC代替奇偶校验），如图16所示，Exascale系统SRAM未校正错误率将只有3-6.5倍Cielos未校正错误率，如图16所示。我们从此分析得出的结论是，供应商应积极主动地降低未校正错误率SRAM故障。然而，由于技术缩减，SER可以通过降低SER来获得大幅度的潜在减排，而大部分剩余部分可能通过对单位故障进行更正校正。然而，一旦达到实际的限制，可能需要更先进的技术来降低多位故障的速率。

11.总结

可靠性在未来几年将继续是一个重大挑战。了解实践中遇到的故障的性质可以使所有利益相关者受益，包括处理器和系统架构师，数据中心运营商，甚至是应用程序的作者，寻求设计更具弹性的大型数据中心和系统。

在本文中，我们介绍了DRAM和SRAM故障的数据，量化了几种硬件恢复技术的影响，并提出了可靠的系统设计方面的经验教训。

我们的研究结果表明，尽管多年来系统取得了长足的进步（例如，从SEC-DED ECC转向DRAM子系统的芯片），但为了为未来的大型企业提供强大的平台，还有更多的工作要做，规模计算系统。