

引用格式:梁伟伟,王晶,王旭,等. 体系结构特征参数指导的近似计算技术[J]. 微电子学与计算机,2019,36(12):72-77. [LIANG W W, WANG J, WANG X, et al. Architecture parameters guided approximate computing technique[J]. Microelectronics & Computer, 2019, 36(12): 72-77.]

体系结构特征参数指导的近似计算技术

梁伟伟^{1,2}, 王 晶^{1,3}, 王 旭^{1,2}, 刘鑫培¹, 张伟功^{1,4}

(1 首都师范大学 信息工程学院, 北京 100048; 2 电子系统可靠性技术北京市重点实验室, 北京 100048;

3 中国科学院 计算技术研究所, 北京, 100080; 4 北京成像理论与技术高精尖创新中心, 北京 100048)

摘 要: 能耗是制约微处理器发展的瓶颈之一, 近似计算技术能够放松对精度的要求提高能效性. 不同的应用程序对精度的需求各不相同, 但现有技术选择近似程度时没有考虑程序需求, 因而无法保证达到最优的能效性. 针对上述问题本文提出了体系结构特征参数指导的近似计算技术, 建立了体系结构参数同近似敏感度之间的联系, 通过机器学习方法选择对近似影响较大的参数, 利用分类决策树建立区分程序近似敏感度的分析模型, 并基于分析结果指导近似程度的选择. 实验结果显示, 本文的方法能够为不同程序选择合适的近似程度, 实现计算精度和能效性的折中优化.

关键词: 近似计算; 体系结构; 能效性; 主成分分析; 分类决策树

中图分类号: TP302

文献标识码: A

文章编号: 1000—7180(2019)12—0072—06

Architecture-parameters-guided approximate computing technique

LIANG Wei-wei^{1,2}, WANG Jing^{1,3}, WANG Xu^{1,2}, LIU Xin-pei¹, ZHANG Wei-gong^{1,4}

(1 College of Information Engineering, Capital Normal University, Beijing 100048, China;

2 Beijing Key Laboratory of Electronic System Reliability and Prognostics, Beijing 100048, China;

3 Institute of Computing Technology, Chinese Academic of Science, Beijing, 100080, China;

4 Beijing Advanced Innovation Center for Imaging Theory and Technology, Beijing 100048, China)

Abstract: This paper proposes an approximate computing technique for the guidance of architectural characteristic parameters. The technique analyzes the relationship between architectural parameters and approximate sensitive features. The parameters with great influence on approximate computing are selected by machine learning method. The classification and regression tree is used to analyze the approximate sensitivity of application and select the approximate degree according to the analysis results. Experimental results show that the method proposed by the paper can select the appropriate approximate degree for different application, and achieve the tradeoff optimization between computational accuracy and energy efficiency.

Key words: approximate technique; architecture; energy efficient; principal components analysis; classification and regression trees

收稿日期: 2019—01—23; 修回日期: 2019—03—01

基金项目: 国家自然科学基金项目(61772350); 共有信息系统装备预先研究项目(公开)(JZX2017—0988/Y300); 北京市科技新星计划(Z181100006218093); 体系结构国家重点实验室开放课题(CARCH201607); 北京未来芯片技术高精尖创新中心科研基金资助项目(KYJJ2018008); 科技创新服务能力建设—基本科研业务费(科研类)(19530050173, 02518530500)

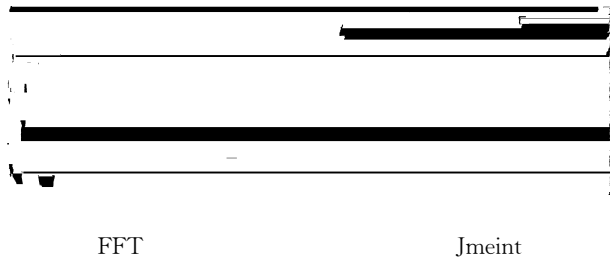
1 引言

随着科学计算、社交媒体和网络游戏等应用的飞速发展和普及,计算机系统运算和存储所消耗的能源也在急剧增加. 如何降低计算机系统的能耗成为体系结构领域研究的重要问题.

现有系统为维持“精确”计算会带来高昂的能耗开销,事实上许多“精确”计算并不是必要的. 例如,由于人本身的感知能力有限,一些视频应用中可以跳过部分帧而不影响人的观赏体验,用户通常只在乎这些应用的结果,不关心中间计算过程是不是“精确”执行. 这些应用输入、输出和计算过程往往本身有天然的容错性,可以通过近似计算放松对精度的要求,大幅提升系统的能效性.

基于近似计算技术良好的能效性,研究人员从不同的层面提出多种实现方案. 电路层可以通过简化电路或者采用不准确的电路器件来实现近似^[1]. 算法层可以通过尾数截断^[2]、基于数据模式的并行处理和基于查表的计算等技术实现近似. 体系结构层面,典型的近似计算技术有基于神经网络的加速器设计,该技术用神经网络代替循环迭代过程,并利用 NPU/GPU 加速神经网络的执行^[3],达到通过神经网络近似提高能效性的目的.

上述近似技术都可以根据方案的特点实现不同近似程度的选择. 例如尾数截断技术,截断的位数越多近似程度越高,相反截掉的位数越少近似程度越低. 基于神经网络的近似方法,网络拓扑越简单近似程度越高,网络拓扑越复杂近似程度越低. 无论哪种技术,近似程度越高,计算结果的误差越大但同时能耗也越低;而近似程度越低,计算结果的误差较小但相应的系统能耗则较高.



FFT Jmeint
图 1 不同程序在不同近似程度下的精度损失

进一步实验发现,随着近似程度的变化,不同的程序在精度损失方面的变化差别很大. 本文采用基于神经网络的近似技术对比了 FFT 和 Jmeint 两个程序中近似程度对输出精度的影响,如图 1 所示. 随着近似程度的增加 FFT 的输出质量急速下降,而

Jmeint 则几乎没有下降. 分析用户期望的计算偏差最小的网络拓扑可以看到,FFT 由于采用精确数值作为输出,因此需要复杂的神经网络才能达到精度要求,属于近似敏感型程序. 而同样实现用户可接受的精度,由于 Jmeint 输出的是空间两个三角形是否相交的统计结果,对神经网络拓扑复杂度的需求则大大降低,属于近似不敏感程序. 近似敏感型程序需要较低的近似程度即复杂的网络拓扑保证输出精度,而对于 Jmeint 这类近似不敏感的程序采用复杂的网络拓扑不仅起不到提高输出精度的作用反而会带来更多的乘加运算而导致更大的能耗开销和时间开销. 通过实验和分析可以看出,近似程度的选择应该考虑应用程序的近似敏感度来确保获得最佳能效性.

然而,现有的近似计算技术例如 Roy 等人提出了一种利用统计方法自动发现程序中可近似数据的 SW 框架^[4];Grigorian 和 Reinman 等为一些计算密集型应用程序设计了近似计算质量监测器,通过监控不同近似技术和近似程度对质量的影响来确定程序运行配置^[5]. 这些近似技术对不同近似敏感度的程序不加以区分,为了追求输出质量会导致额外的能耗和时间开销. 针对上述问题本文提出了体系结构特征参数指导的近似计算技术,利用机器学习算法分析了能够区分程序近似敏感特征的体系结构参数,建立了基于近似相关的体系结构参数的近似程度选择模型,实现了既能保证输出精度又可以降低能耗和时间开销的近似计算技术.

2 基于体系结构参数的近似计算技术

程序运行过程中产生的体系结构参数反映了程序的特征,例如在计算指令和访存指令的比例方面,FFT 应用计算操作所占的比例大于访存操作,而 Jmeint 应用在大量数据上执行相同的操作,因此访存比例与计算比例相当. 这些体系结构特征恰恰反映了近似的特征,对于 FFT 应用的大量数值计算引入近似会导致计算结果的较大误差,而对于访存密集型的 Jmeint 输入本身就会存在一定的噪声和冗余,并且大量数据的相似操作对计算结果的精度变化不敏感,这同实验发现的 FFT 近似敏感和 Jmeint 近似不敏感的特性一致. 基于这个发现,本文选取能够反映程序特征的体系结构参数,建立近似程度、输出精度同体系结构参数之间的关联关系.

体系结构特征参数指导的近似计算技术流程如图 2 所示,首先选取不同类型的代表性应用程序,运

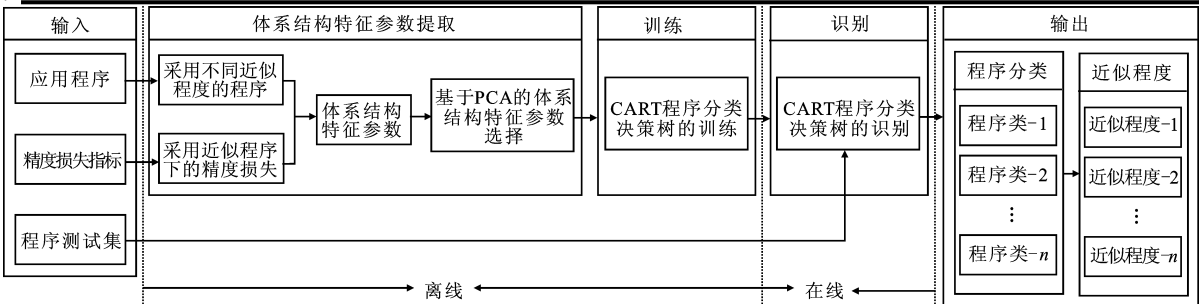


图 2 体系结构特征参数指导的近似计算技术流程

行程序并收集不同程序在不同近似程度下的输出精度损失和体系结构参数. 通过主成分分析技术筛选出对输出精度影响较大的参数, 同时达到对特征参数降维的效果. 将选取出的特征参数和精度变化特征作为输入, 训练分类决策树, 使得具有“类似”近似敏感特性的程序分为一个“程序类”, 同一个“程序类”中达到最优能效比的近似程度选择也相同. 通过分类决策树的分析和典型程序的能效性与近似程度的分析, 实现基于特征参数分类的程序近似程度选择指导方案.

2.1 体系结构特征参数选取

程序在动态运行中会产生大量的体系结构特征参数, 我们需要从中提取能够刻画程序近似敏感特性的参数, 但是人工手动的分析时间开销和人力成本都比较高, 同时也无法保证结果的准确性和可靠性. 为此本文提出了一种基于机器学习的体系结构特征参数选择技术, 支持自动化的判定程序的近似敏感度.

2.1.1 近似角度的特征参数分类

指令执行的数据通路相似的情况下, 其受近似影响所产生的输出质量变化也相似^[6], 也就是指令对近似有相同的敏感度. 而指令执行的特征通常反映在体系结构参数上, 因此相似数据通路上产生的体系结构参数也会有相似的近似敏感特性. 分析体系结构参数在控制路径上的行为特征可以有效地反映程序的近似敏感特性, 因此从受近似影响的角度, 将体系结构参数分为三大类:

第一类从指令类型角度选取的特征参数: 程序的某些指令因近似发生改变后不会对输出质量产生任何的影响: 1 例如 NOP 指令是空语句, 不做任何操作, 近似导致的变化不会对输出结果有任何影响, 因此该指令产生的体系结构参数对近似不敏感. 2 分支预测指令产生的体系结构参数对近似也不敏感, 指令基于预测结果执行后并不立刻提交, 而是在

顺序执行确认分支执行结果正确之后再提交执行结果, 因此近似产生的分支预测误差不会影响程序最终运行. 3 流水线停顿指令^[7]只影响系统性能, 而不影响精度, 因此是近似不敏感指令. 其对应的体系结构参数, 如取指造成的停顿占总周期的比例、寄存器分配表造成的停顿占总周期的比例和资源相关停顿占总周期的比例等可以代表近似不敏感程序的特征.

第二类是存储相关的体系结构特征参数. 以近似不敏感的图像应用为例, 大量像素数据的读取使得程序访存指令比例较高, 存储相关的体系结构特征参数同包含大量精确运算的计算密集型程序相比差异较大. 由此可见访存能够反映程序的近似敏感特性. 访存相关的体系结构特征参数分为: 1 访存指令比例: 读写操作比例、核心态指令比例和用户态指令比例. 2 缓存行为参数: 每千条指令 L1/L2/L3 缓存缺失率、每千条指令中 load 操作命中 L2 cache 次数/命中其他核心的 L2 cache 次数/命中 L3 cache 次数等^[8]. 3 一致性监听协议参数: 每千条指令中监听协议命中次数/独占命中次数/修改命中次数. 4 访存并行度参数 MLP.

第三类是反映计算操作相关特性和比例的参数, 通常含有大量数值计算和浮点计算的程序对近似比较敏感, 例如科学计算中精确的数值运算一旦出错会造成灾难性的后果. 这类特征分为描述计算特征和计算访存比例特征的参数: 1 计算相关特征参数: 整型操作比例和浮点操作比例. 2 计算访存比特征: 整型计算与访存的比例和浮点计算与访存的比例.

2.1.2 基于主成分分析的特征参数选择

上一节所分类和选取的体系结构特征参数反映了程序的特征, 但并不是所有参数都跟程序近似敏感度有关联. 此外, 提取的体系结构参数过多会导致基于体系结构参数的程序分类过程时间过长, 而某些体

系结构特征参数之间存在着相关性,例如长延迟的缓存缺失可能会引起处理器流水线的停顿.因此所提取的参数也存在一定的冗余.如何找到能够反映程序近似特征的参数,并且选取最小集是我们解决的关键问题.为此,本文选取主成分分析算法,分析选取对近似敏感度影响较大的参数,同时分析参数之间的相关性消除冗余参数从而降低参数的维度.

主成分分析(Principal Component Analysis, PCA),是一种广泛用于数据挖掘的机器学习算法,能够把多指标转化为少数几个综合指标.将前文选取的三大类与近似相关的体系结构特征参数通过主成分分析之后选取出 12 个特征贡献率高于 95% 的参数用于刻画程序的近似敏感特性,结果如表 1 所示.其中包括存储类,转移预测类等.一致性监听协议相关的指令由于同近似关联不大,因此没有被选取.并且访存类参数中存在相关冗余的参数也被剔除,例如 L2 MISS 与 L2 HIT 具有一定的相关性,PCA 只选取 L2 MISS 而不选 L2 HIT 确保了所选的体系结构特征参数没有潜在的相关性.综上所述,通过 PCA 选取的体系结构特征参数可以很好的反映程序的近似敏感特性,同时又可以消除体系结构特征参数之间的相关性.

表 1 体系结构特征参数的选取结果

类别	体系结构特征参数	含义	对近似的敏感程度
指令类型角度特征参数	BRANCH	分支操作指令比例	受近似的 影响很小
	BR MISS	分支预测错误率	
	FETCH STALL	取指造成的停顿占总周期的比例	
存储操作特征参数	MLP	访存并行度	受近似的 影响较小
	L2 MISS	每千条指令中 L2 cache 缺失次数	
	L3 MISS	每千条指令中 L3 cache 缺失次数	
	L1I MISS	每千条指令 L1 指令 cache 缺失次数	
计算访存比例参数	SSE FP	SSE 浮点操作比例	受近似的 影响较大
	FP TO MEN	浮点计算和访存的比例	
	ILP	指令并行度	
	UOPS EXE CYCLE	每个周期内执行的微操作数	
	FP	浮点操作比例	

2.2 基于程序分类的近似程度选取

程序的近似程度选择直接影响着系统的输出精度和能效性,建立体系结构参数同近似敏感度之间

的关联是分析和预测程序的近似敏感程度的关键.针对上述需求,本文选择了机器学习经典算法 CART 分类决策树自动化的判定程序的近似敏感特性.我们通过 CART 算法将已知敏感度的测试集中程序动态运行获得的体系结构特征参数构成一个数据集,采用二分递归的分裂技术,利用不断的循环迭代将训练数据集分为两个子样本集分别代表不同的近似敏感度.

将目标程序的体系结构特征参数输入到分类决策树之后程序被分到具有不同敏感度的程序类中,测试集中对于程序敏感度的分类根据不同程序的功能和结果形式划分.一种程序类是近似敏感型程序,以典型的金融数据分析程序 Blackscholes 为例,计算过程主要完成复杂的算法流程,其中包含大量的精确数值计算,输出精度主要通过精确计算的数值结果和近似计算的数值结果之间的差距来评估,对于这类程序数值的微小变化都会导致输出精度的损失,因此近似敏感度高,为了保证精度往往选择较低的近似程度.

另一类程序是近似不敏感程序,日常应用中存在大量的程序对近似不敏感,例如 JPEG 图像处理算法,由于人眼感知的限制输出精度损失只需要统计像素值偏差超过某一阈值的像素点的比例,像素点产生的轻微偏差对输出精度影响不大,因此对于这类程序可以利用程序自身较高的容错性选择较高的近似程度,进一步降低系统能耗,提高能效性.

在分类决策的第一层可以将程序根据上述分析分为近似敏感和不敏感,但在这两种典型程序之间,还有部分程序的输出精度与近似程度之间既不是正比例的变化,也不是保持不变,而是输出精度随着近似程度的减小先增大后保持不变.典型的程序是基于概率和统计的机器分类和识别算法,通常包含大量的矩阵计算,其输出精度的计算是比较精确计算的结果和近似计算结果的 L2 范数.对于这类程序近似程度和能效性之间通常存在一个拐点,拐点之前随着近似程度的降低精度提高,拐点之后进一步降低近似程度不会再引发较大的精度提升,因此为提高能效性带来了更大的优化空间.为了找到拐点,需要决策树继续分类,以获得准确的近似敏感度,指导近似程度的合理选择.

3 实验结果及分析

为了分析本文所提出的方案对于程序近似敏感度分析和近似程度选择的效果,实验选择了经典的

近似计算基准测试程序集 Axbench, Axbench 提供了针对不同领域应用程序特征的精度损失指标. 实验采用基于神经网络代替循环迭代过程的近似技术, 通过改变神经网络的拓扑复杂度实现对不同近似程度的支持, 利用网络的均方误差(mse)来量化神经网络的近似程度.

表 2 程序的分类以及网络选取结果

分类	程序	层数	每层神经元数
程序类-1	Blackscholes	2	64-64
程序类-2	Jmeint、Jpeg、Sobel	1	2
程序类-3	Inversek2j、FFT	1	16

对于典型应用程序近似敏感分类结果和近似程度选择结果如表 2 中所示. 为了验证分类结果是否可以准确的预测程序的近似敏感特性和选取的近似程度的是否合理. 我们评估了表 2 中 6 个程序在不同的 mse 即网络拓扑下运行的输出精度损失, 其结果如图 3 所示. 其中横轴是表示近似程度的 mse, 纵轴是程序近似之后的精度损失, 红圈标注出来的是通过本文方法选出的近似程度在曲线上的位置. 观察图 3 可以发现图像应用 Jpeg 和 Sobel, 以及概率运算应用 Jmeint 分为了“程序类-2”, 这类程序本身就具有天然的容错特性对近似不敏感, 所以他们的近似敏感特性曲线几乎不随网络拓扑的改变而改变, 针对此类程序, 本文选取的网络拓扑为隐藏层层数为 1 层, 隐藏层神经元个数为 2, 例如程序 Jmeint 在该网络配置下的对应的近似程度即 mse 为 0.05, 该网络配置下的近似程度很高同时精度几乎没有任何损失; 程序分类方案将具有大量数值运算的 Blackscholes 分类到“程序类-1”, 这类程序容错性不好对近似比较敏感, 输出精度会随着近似程度的增大而一直减小, 针对此类程序本文选取的网络拓扑为隐藏层层数为 2 层, 每层隐藏层的神经元个数都为 64, 该网络配置下对应的近似程度即 mse 为 2.3×10^{-5} , 可以使得精度损失较小. 含有大量矩阵运算的 Inversek2j 和 FFT 被分到“程序类-3”, 这类程序, 输出精度随着近似程度的减小先增大后保持不变, 本文选取的网络拓扑为隐藏层层数为 1 层, 隐藏层神经元个数为 16, 例如 Inversek2j 在该网络配置下对应的 mse 为 0.01, 选择的网络拓扑对应的近似程度都是程序近似敏感特性曲线上的“拐点处”, 这样既可以保证精度损失最小又可以避免由于近似程度减小而选用复杂的网络带来额外的能耗开销. 综上所述, 实际测试得到的各个程序的近似敏感特性曲线与表 2 通过 CART 分类决策树给出的分类

完全吻合, 同时选出的近似程度符合程序近似敏感特性的要求.

支持近似的计算模式中, 除了传统的能耗和延迟, 输出精度也是必须关注的指标, 单一指标的优化无法满足用户多维度的需求, 例如采用非常简单的网络拓扑来近似能耗和时间都非常低, 但输出精度却有可能降得太低无法满足用户的需求. 为了进一步的评估近似程度选择方案对能效性的改善, 我们采用了综合应用程序输出精度、能耗和时间的指标 EDQ:

$$EDQ = \frac{Energy \cdot Delay}{Quality}$$

(1)

式中, *Energy* 为能耗, *Delay* 为延迟时间, *Quality* 为输出精度. 指标中能耗和延迟越小越好, 输出精度越大越好, 所以 EDQ 越小越好.

我们对比了本文方案和基于精度的近似程度选择方案^[3]的综合优化效果, 如图 4 所示. 结果显示本文方案选择的近似程度考虑了能耗, 整体能效性优于现有方案, 特别是 Inversek2j 按照本文选取的网络拓扑进行近似计算比现有方案 EDQ 值下降了将近 70%. 上述实验证明本文提出的体系结构特征参数指导的近似计算技术选择的网络拓扑可综合权衡能耗、延迟和输出精度, 达到综合优化的效果.

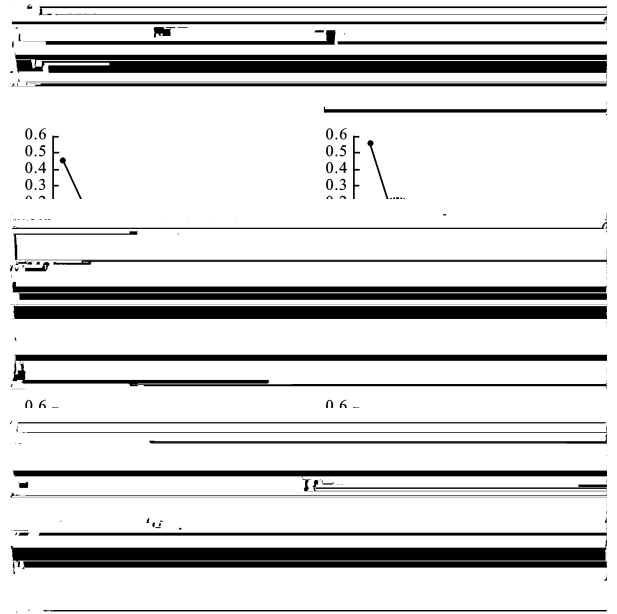


图 3 各程序在不同近似程度下的精度损失

4 结束语

近似计算技术利用程序的容错性, 为提高能效性带来契机. 分析发现程序运行产生的体系结构特

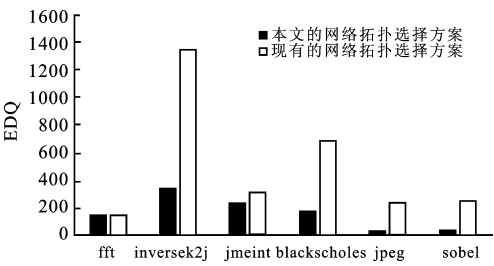


图 4 网络拓扑选择方案对比

征参数可以有效反映程序的近似敏感特性,本文提出了以体系结构特征参数为指导的近似计算技术,通过机器学习算法分析程序的体系结构特征参数来判定其近似敏感特性,并根据程序的近似敏感特性给出最佳近似程度.实验结果证明本文提出的方案能够在保证输出精度的同时优化系统能效性.

参考文献:

[1] KAHNG A B, KANG S. Accuracy-configurable adder for approximate arithmetic designs[C]// Proceedings of the 49th Annual Design Automation Conference, USA, California; ACM, 2012; 820-825.

[2] HASHEMI S, ANTHONY N, TANN H, et al. Understanding the impact of precision quantization on the accuracy and energy of neural networks[C]//Proceedings of the Conference on Design, Automation & Test in Europe, Switzerland, Lausanne; EDAA, 2017: 1478-1483.

[3] ESMAEILZADEH H, SAMPSON A, CEZE L, et al. Neural acceleration for general-Purpose approximate programs[J]. IEEE Micro, 2013, 33(3):16-27.

[4] CHIPPA V K, CHAKRADHAR S T, ROY K, et al. Analysis and characterization of inherent application resilience for approximate computing[C]// Proceedings of the 50th Annual Design Automation Conference, USA, Texas; ACM, 2013;1-9.

[5] GRIGORIAN B, REINMAN G. Dynamically adaptive and reliable approximate computing using light-weight error analysis[C]// Adaptive Hardware and Systems, UK, Leicester; IEEE, 2014;248-255.

[6] VENKATAGIRI R, MAHMOUD A, HARI S K S, et al. Approxilyzer: Towards a systematic framework for instruction-level approximate computing and its application to hardware resiliency[C]// The 49th Annual IEEE/ACM International Symposium on Microarchitecture, Taiwan, Taipei;IEEE, 2016;1-14.

[7] 张迅珍, 梁青, 李涛. RV32I 控制单元设计与实现[J]. 微电子学与计算机,2018, 35(3):74-78.

[8] 王珍珍,杨志涛,王晶,等. 基于 BCH 算法的高速缓存纠错检错方案研究[J]. 微电子学与计算机,2016, 33(11):124-128.

作者简介:

梁伟伟 男,(1992-),硕士研究生.研究方向为计算机体系结构.

王 晶(通讯作者) 女,(1982-),副研究员,硕士生导师.研究方向为计算机体系结构. E-mail:jwang@cnu.edu.cn

王 旭 男,(1993-),硕士研究生.研究方向为计算机体系结构.

刘鑫培 男,(1995-),学士.专业方向为计算机科学与技术.

张伟功 男,(1967-),研究员,博士生导师.研究方向为高可靠嵌入式系统.

(上接第 71 页)

参考文献:

[1] 方晓斌. 具有快速瞬态响应的电压模式 Buck 变换器设计[D]. 合肥:合肥工业大学,2016.

[2] 解光军,程心. DC/DC 开关变换器的建模与设计[M]. 北京:科学出版社,2015.

[3] 杨令. DC-DC 变换器驱动优化设计与稳定性研究[D]. 西安:西安电子科技大学,2015.

[4] LIU J M, WANG P Y, KUO T H. A Current-Mode DC-DCBuckConverter with Efficiency-Optimized Frequency Control and Reconfigurable Compensation[J]. IEEE Trans Power Electron,2012,27(2):869-880.

[5] XIAO J W, ANGEL V PETERCHEV, ZHANG J H, et al. A 4-ua quiescent-current dual-mode digitally controlled buck converter IC for cellular phone applications[J]. IEEE J Solid-State Circuits, 2004, 39(1): 2342-2348.

作者简介:

方贤朋 男,(1994-),硕士研究生.研究方向为微电子与集成电路设计.

王兴华(通讯作者) 女,(1983-),讲师,硕士生导师.研究方向为微电子与集成电路. E-mail:89811@bit.edu.cn