

多尺度样本熵在流量矩阵分析与评估中的应用

颜若愚

(河南财经政法大学 计算机与信息工程学院, 河南 郑州 450002)

摘 要: 针对网络用户行为的特点提出利用多尺度样本熵的两个评价规则来分析 P-P flow 流量矩阵, 该方法能够对网络流量表现出来的一些特征进行合理解释, 并得到一些有趣的结论. 随后在多尺度样本熵分析特性基础上提出一种评估流量时间序列健康度的方法, 通过对真实网络流量和仿真流量的实验, 表明该方法简单有效, 能够准确合理评估网络流量的健康程度.

关键词: 多尺度分析; 样本熵; 异常检测; 流量分析; 流量矩阵

中图分类号: TP393.08

文献标识码: A

文章编号: 1000-7180(2014)03-0056-05

Applying Multi-scale Sample Entropy to Analyze and Evaluate Traffic Matrix

YAN Ruo-yu

(College of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou 450002, China)

Abstract: For the features of network user behavior, two evaluation rules of multi-scale sample entropy are used to analyze P-P flow traffic matrix. The method can reasonably explain some of the features shown in the network traffic and draw some interesting conclusions. Then a method to evaluate healthy degree of traffic time series is proposed based on analysis capacity of multi-scale sample entropy. Experiments conducted with actual network traffic and simulation network traffic show that the proposed method is simple and effective, and can accurately evaluate the healthy degree of network traffic.

Key words: multi-scale analysis; sample entropy; anomaly detection; traffic analysis; traffic matrix

1 引言

目前基于信息理论度量的方法在网络安全领域中得到了广泛研究和应用, 最常见的信息理论, 比如香农熵^[1]、雷尼信息熵等, 能够简单而有效地分析网络流量及其异常状况. 虽然信息理论度量方法的有效性毋庸置疑, 但它们都只在单时间尺度上针对网络流量进行度量, 并不能在多时间和空间尺度上揭露网络流量的一些复杂行为和特性. 为此, 需要寻找一种能够衡量时间序列在多尺度上的复杂度方法, 多尺度样本熵就是这样一种方法, 该方法侧重于在

多时间尺度上量化网络流量动力学所表达的信息. 与传统的复杂度测量方法相比, 它具有可应用于有限长度时间序列的优点; 而与传统的多尺度分析方法, 如自相似性^[2]、小波分析^[3]等相比, 它具有算法复杂度低, 简单易用的优点.

多尺度样本熵分析首先由 Madalena Costa 等人提出并用于研究分析生物信号如脑电图、心电图等^[4], 实验表明该方法在分析和检测异常生物信号时具有很好的效果. 本文将多尺度样本熵应用到网络流量矩阵的分析以及流量健康度评估中. 流量矩阵反映的是一个网络中所有源节点和目的节点对之

收稿日期: 2013-05-20; 修回日期: 2013-07-10

基金项目: 国家自然科学基金项目(61101211, 61202285); 河南省自然科学基金项目(132300410337); 河南省教育厅项目(13B520901)

间的流量交换关系,其中 OD flow 流量矩阵是一种 PoP 级的流量矩阵,P-P flow 流量矩阵则反映一个路由器中所有活动端口之间的流量交换关系,属于端口级的流量矩阵.本文以 OD flow 流量矩阵和 P-P flow 流量矩阵作为研究对象,提出了分析和评估流量矩阵的方法,通过对真实流量和仿真流量的实验,表明多尺度熵方法是有效的.

2 样本熵和多尺度样本熵

2.1 样本熵

样本熵实质是 Renyi 广义熵在阶 $\alpha = 2$ 上的近似,其特点为估计结果是无偏的.假设 $\{X_i\} = \{x_1, \dots, x_i, \dots, x_N\}$ 是一个长度为 N 的时间序列,则其中维数为 m 的所有子序列向量表示为: $u_m(k) = \{x_k, x_{k+1}, \dots, x_{k+m-1}\}, 1 \leq k \leq N-m$. 用 $n_i^m(r)$ 表示向量 $u_m(i)$ 和其他向量 $u_m(j)$ 近似的次数,即有多少个向量 $u_m(j)$ 满足 $d[u_m(i), u_m(j)] \leq r$, 在这里 $d[u_m(i), u_m(j)] = \max\{|x_{(i+l)} - x_{(j+l)}| : 0 \leq l \leq m-1\}$, r 为一个正常数.

注意 $u_m(i)$ 并不和自身进行比较,并且关于一个有 N 个数据的时间序列, $u_m(i)$ 只考虑和前 $N-m$ 个长度为 m 的向量做比较,以保证对于 $1 \leq k \leq N-m$, 长度为 $m+1$ 的向量 $u_{m+1}(k)$ 也能够有定义.

等式 $U_i^m(r) = n_i^m(r)/(N-m)$ 即表示向量 $u_m(i)$ 和任意其他向量 $u_m(j)$ 相似的概率. 则 $U_i^m(r) (1 \leq i \leq N-m)$ 的均值为

$$U^m(r) = 1/(N-m) \sum_{i=1}^{N-m} U_i^m(r),$$

表示任意两个向量在距离 r 内相似的概率. 于是样本熵定义为

$$S_E(m, r) = \lim_{N \rightarrow \infty} -\ln \frac{U^{m+1}(r)}{U^m(r)} \quad (1)$$

其统计量的估计形式即为

$$S_E(m, r, N) = -\ln \frac{U^{m+1}(r)}{U^m(r)} = \ln \frac{\sum_{i=1}^{N-m} n_i^m(r)}{\sum_{i=1}^{N-m} n_i^{m+1}(r)} \quad (2)$$

S_E 正好为条件概率的自然对数的负值,该条件概率表示如果长度为 m 的序列相互近似,那么当在每个序列中增加一个新数据时得到的新序列同样会相互近似. 样本熵度量了一个时间序列的规则性程度,当熵值越大,表明时间序列越不规则,即越复杂,熵值越小,表明时间序列的规则性越好.

2.2 多尺度样本熵

为了按不同时间尺度因子 τ 构造一系列的粗粒度时间序列,首先将原始的一维离散时间序列 $\{X_i\} = \{x_1, \dots, x_i, \dots, x_N\}$ 依次划分为不重叠的长度为 τ 的多个子序列,然后计算每个子序列内数据的平均值,则粒度为 τ 的时间序列都可以根据以下公式计算得到.

$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i, 1 \leq j \leq N/\tau \quad (3)$$

于是粒度为 τ 的时间序列的长度等于原始时间序列除以尺度因子 τ 的长度. 当尺度因子为 1 时,时间序列 $\{y^{(1)}\}$ 即为原始时间序列. 最后计算每个粗粒度时间序列的样本熵,以尺度因子 τ 为横坐标,以对应的样本熵为纵坐标绘制成图的过程就是所谓的多尺度样本熵分析. 本文就是利用多尺度样本熵的一些特点分别对矩阵流量进行分析和健康度评估.

3 P-P flow 矩阵流量的采集及其特征分析

本文使用的原始流量数据为校园网内一个 10 端口路由器近一天(时间为 2007.12.27 4:01 PM ~ 2007.12.28 2:36 PM)的 Netflow 流量,采样时间为 1 分钟. 利用 Netflow 记录中 input 和 output 属性对 Netflow 流量进行汇聚,统计得到 P-P flow 流量,本文定义的 P-P flow 流量即单位时间内从同一个端口进入路由器并从同一个端口出去的所有数据包/字节/Netflow 流的集合(即 Port to Port). 图 1 显示了以网络流数为度量的 4 个 P-P flow 流量样本,它们的目的端口都为访问外网的第 7 号端口.

由于采集 Netflow 流量的路由器部署在学生生活区,不同类型学生的生活学习特点会在这些 P-P flow 流量样本中有所体现. 比如在时刻点 500(实际时间为午夜 0:21 AM)附近,P-P flow 1-7,5-7 和 8-7 流量出现一个最高峰,然而 P-P flow 4-7 出现流量递减的趋势. 这主要是因为前 3 个 P-P flow 流量样本来自于研究生生活区,而后一个 P-P flow 流量样本来自于本科生生活区. 研究生白天一般都在实验室度过,因此宿舍区的网络流量不太大,但他们往往有午夜还在宿舍上网查资料做研究的习惯,导致午夜这段时间网络流量最大. 然而本科生的生活和学习方式更有规律,晚上学校会定时对宿舍拉闸断电,因此午夜至凌晨这段时间基本上没有什么流量出现.

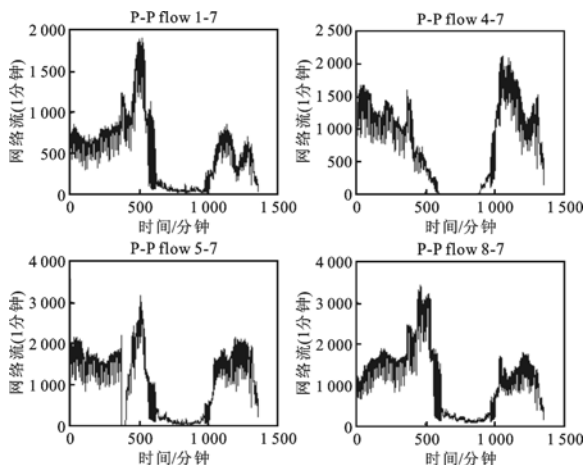


图1 流向第7号端口的4个P-P flow流量的样本

4 多尺度样本熵分析 P-P flow 流量矩阵

4.1 分析方法

使用多尺度样本熵分析 P-P flow 流量矩阵参考了如下两个规则:(1)一个时间序列如果在大多数时间尺度上熵值高于另一个时间序列,那么前者被认为比后者更复杂;(2)如果熵值随着时间尺度增大而单调递减,表明原始信号在最小时间尺度上包含最多的信息量,即复杂度较大,规则性较差^[4]。

为了提高多尺度样本熵的分析和评估性能,在分析原始数据前有必要对其进行预处理,即使用最大最小线性变换方法来归一化原始序列: $z_i = (x_i - x_{\min}) / (x_{\max} - x_{\min})$,其中 $x_{\min}$ 为原始序列中最小值, $x_{\max}$ 为原始序列中最大值, z_i 为对原始数据 x_i 归一化后的结果.若无特别说明,本文只计算尺度为1~20的样本熵以保证计算精度,且计算样本熵的子序列维数为 $m=2$,向量距离值固定为 $r=2$ 。

4.2 网络流数度量下的 P-P flow 流量样本分析

图2是对图1中P-P flow流量样本进行多尺度样本熵分析后的结果.在所有时间尺度上,P-P flow 5-7流量样本的样本熵高于P-P flow 1-7,4-7和8-7时间序列的样本熵,由多尺度样本熵分析的规则1可知,前者比后者具有更复杂的结构.究其原因是在374~403(10:15 PM~10:44 PM)期间网络设备发生了故障,导致P-P flow 5-7样本在该段时间的流量为0,使得该样本的形状变得更复杂.这一分析结果表明多尺度样本熵分析可用于评估流量样本的健康程度.此外,由图可见P-P flow 1-7和8-7流量样本的多尺度样本熵值很接近,并且在所有时间尺度上几乎具有相同的趋势,这也意味着它们在所有时间尺度上具有相似的形状.然而P-P

flow 4-7的熵值在所有时间尺度上始终高于P-P flow 1-7和8-7的熵值,这种现象的根本原因可能是由研究生与本科生在生活学习习惯上存在的差异造成的。

进一步分析图2中各P-P flow样本的多尺度样本熵发现,尺度因子为1的熵值并不是所有尺度中熵值最小的,熵值最小时的尺度因子在4左右,然后随着尺度因子的增大熵值基本呈递增趋势.根据多尺度样本熵分析的定义及规则2可知,尺度因子为4左右时流量样本的复杂度最小,最具有规则性,这意味着在子序列长度为 $m=2$ 时,P-P flow样本中长度为5或6的子序列具有最相似的结构,这也间接表明了目前对路由器中Netflow流量采集的时间间隔大多数设为5分钟的合理性,因为该采样时间间隔下流量更具有规则性,更容易突出异常流量.图3所示为当子序列长度 m 为5时进行多尺度样本熵分析的结果.该图结果和图2的结果高度一致,显然尺度因子为1时熵值最小,表明子序列长度为5或6时相似性最高。

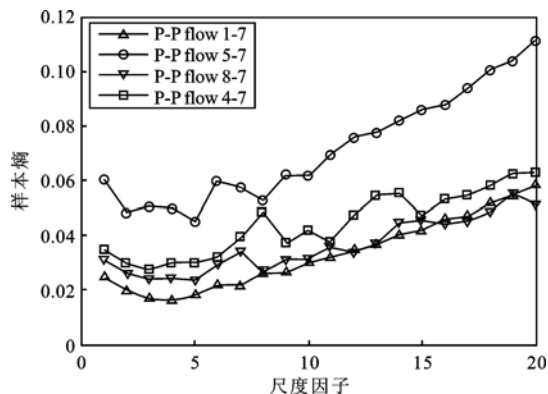


图2 $m=2$ 时以网络流数为度量的 P-P flow 样本的多尺度样本熵分析

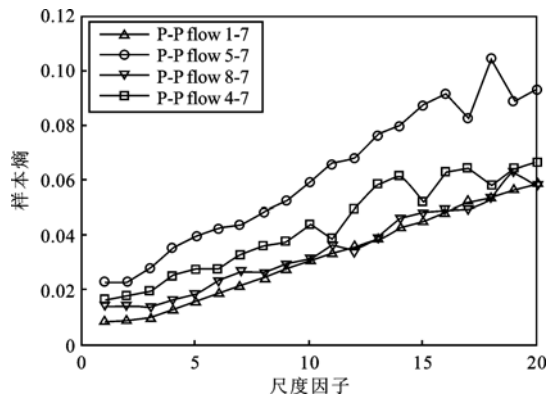


图3 $m=5$ 时以网络流数为度量的 P-P flow 流量样本的多尺度样本熵分析

4.3 三种流量度量下的 P-P flow 流量样本分析

图4中显示了分别以字节数、数据包数、网络流数为度量单位的 P-P flow 8-7 流量样本的多尺度样本熵分析结果. 对于字节数和数据包数度量, 随着尺度因子的增大样本熵几乎在所有尺度上递减, 根据多尺度样本熵分析的规则2, 表明 P-P flow 8-7 原始流量包含最多的信息量, 往往不容易发现流量中的异常. 对于网络流数度量, 随着尺度因子的增大样本熵几乎在所有尺度上递增, 而且网络流数度量下的熵值在所有尺度上都低于字节数和数据包数度量下的熵值. 图5中显示了所有 P-P flow 流量的多尺度样本熵均值分析, 和图4相比, 以字节数和数据包数为度量的熵值更加接近, 但又都远大于以网络流为度量的熵值. 进一步检查所有采集到的 P-P flow 流量样本, 发现这些样本几乎都具有这种多尺度特性, 显然这种现象具有普遍性. 基于以上分析, 根据多尺度样本熵分析的规则1可见字节数和数据包数度量下的流量序列更复杂, 这会导致发现流量异常更加困难. 本文认为虽然每种度量在流量异常检测上都很重要, 但网络流数度量比其他两种度量具有更强的异常检测能力, 数据包数度量的检测能力次之. 这一分析结果和重要文献[5]中的实验结果是一致的.

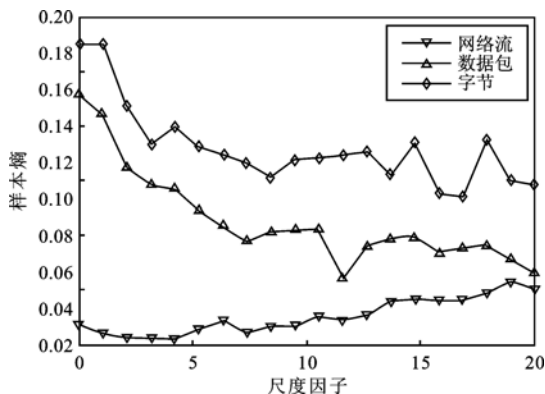


图4 不同度量单位的所有 P-P flow 8-7 流量样本的多尺度样本熵分析

5 多尺度样本熵评估网络流量健康度

5.1 评估方法

多尺度样本熵不仅可以分析网络流量, 而且也可用于评估流量时间序列整体上的健康程度. 该评估方法首先对所有历史正常流量序列进行多尺度样本熵分析, 然后在各尺度因子 i ($i = 1, 2, \dots, n$) 上分别求熵值的均值 $\bar{x}_i$ 和标准差 σ_i , 则流量序列的健康区域就是各尺度因子上上下阈值 $\bar{x}_i \pm 3\sigma_i$ 所确定范围以内的区域. 最后使用多尺度样本熵分析被检

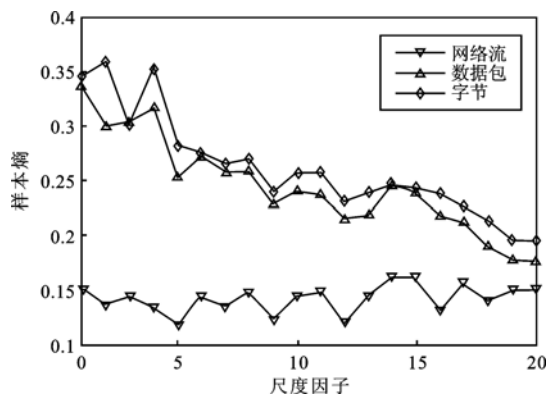


图5 不同度量单位的所有 P-P flow 流量样本的多尺度样本熵均值分析

测流量序列, 如果所有时间尺度上熵值都在健康区域以内, 则认为该流量序列是正常的, 否则有理由怀疑该流量序列的健康状况. 本文提出一个健康度概念来表示被评估流量序列的健康状况, 其计算公式为: 健康度 = 所有尺度因子上熵值在健康区域内的次数 / 尺度因子数. 该式表明被评估流量序列的多尺度样本熵值在健康区域内次数越多, 则越健康.

5.2 评估仿真 P-P flow 流量的健康度

本文选择图1右下子图显示的 P-P flow 8-7 实际流量作为原始流量, 使用文献[6]中介绍的方法生成50个正常流量时间序列和3个分别含有一个异常的异常流量序列. 这三个异常流量序列的异常流量强度依次为0.2、0.5、0.7, 异常持续时间依次为9、15、12分钟.

图6是使用多尺度样本熵分析三个不同异常流量序列的结果, 图中实线是流量健康区域的上下界限, 由50个正常流量序列计算所得, 三个异常流量序列的多尺度样本熵分析都以不同的标记加以区别. 图中异常流量强度分别为0.5和0.7的两个异常流量序列的样本熵几乎都落在健康区域以外, 而异常流量强度为0.2的异常流量序列却没有那么明显. 根据多尺度样本熵异常评估方法, 可计算得到被评估的三个异常流量序列的健康度值按异常流量强度0.2、0.5、0.7的顺序分别为: 0.4、0.2、0.1. 评估结果显示异常流量序列的异常流量强度越大, 异常的持续时间越长, 所计算的多尺度样本熵值通常就越高, 其健康度也就越低, 也越容易被识别为异常.

5.3 评估真实 Abilene OD flow 流量矩阵的健康度

该实验以 Abilene OD flow 公共流量数据集^[7]中共17周的流量(2004-05-01~2004-09-10)作为实验数据, 针对每一天的 OD flow 流量进行健康度

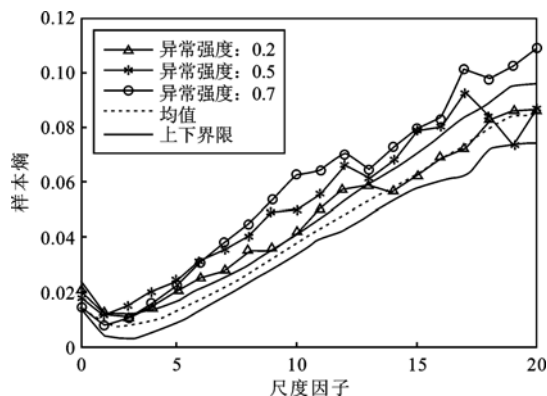


图6 多尺度样本熵分析仿真 P-P flow 异常流量序列评估. 由于不知道数据集中每天 OD flow 流量是否有异常存在, 因此暂且假设它们都是健康的, 从而可使用前面介绍的多尺度样本熵评估流量矩阵异常的方法计算健康区域, 然后根据该健康区域计算每天 OD flow 流量的健康度. 在计算健康区域时考虑到每周中不同工作日的流量特性可能不一样, 所以每一个 OD flow 流量都会针对每周中不同工作日计算各自的健康区域, 即每个 OD flow 流量可分析得到 7 个健康区域图并分别对应一周中 7 天. 由于 OD flow 流量的统计时间间隔为 5 分钟, 一天的数据较少, 为了确保计算精度, 选择多尺度样本熵分析的尺度因子大小为 1~10.

图 7 中显示了第 2 号 OD flow 流量星期六的健康区域和 17 周中每个周六流量的多尺度样本熵分析, 图中较粗的无标记线为健康区域的上下界限. 由图可见所有 17 周的周六流量的多尺度样本熵分析都在健康区域以内, 因此可以认为第 2 号 OD flow 流量在 2004-05-01~2004-09-10 时间内的每个周六流量都是健康的.

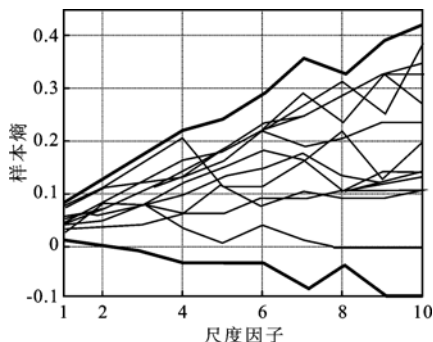


图7 OD flow #2 在周六的健康区域及其 17 个周六流量的多尺度样本熵分析

图 8 显示了第 2 号 OD flow 流量星期五的健康区域和 17 周中每个周五流量的多尺度样本熵分析, 图中较粗的无标记线为健康区域的上下界限. 由图可

见 17 周中绝大部分的周五流量的多尺度样本熵分析都在健康区域以内, 但第 11 周的周五流量的多尺度样本熵分析结果明显偏离健康区域较远, 可计算得其健康度为 0.2, 因此很有可能存在异常情况. 图 9(a) 显示的是第 2 号 OD flow 第 11 周周五流量时间序列, 和图 9(b) 中其平均周五流量序列相比较不难看出, 第 11 周周五流量在 65 时刻左右流量急剧增大是导致其多尺度样本熵增大的主要原因, 而增大的流量很可能就是网络攻击行为所致.

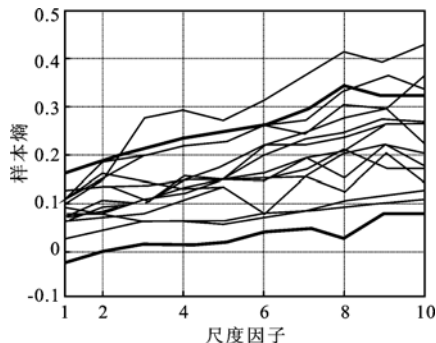


图8 OD flow #2 在周五的健康区域及其 17 个周五流量的多尺度样本熵分析

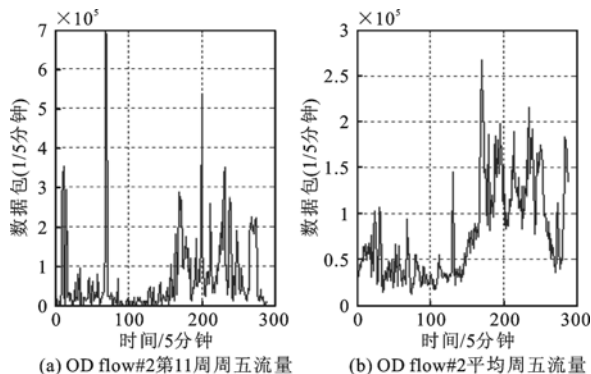


图9 OD flow #2 第 11 周周五流量和平均周五流量的对比

6 结束语

本文提出使用多尺度样本熵来分析流量矩阵以及评估流量时间序列的健康度. 多尺度样本熵的多尺度特性能够很好地刻画不同类型用户的网络使用习惯, 解释路由器中 Netflow 流量采集时间间隔设为 5 分钟的合理性, 并通过对以字节数、数据包数、网络流数为度量的 P-P flow 矩阵流量的多尺度样本熵分析说明了为什么网络流数度量比其他两种度量具有更强的异常检测能力. 最后在多尺度样本熵的分析特性基础上提出一种基于多尺度样本熵的流量健康度评估方法, 该方法能够准确合理评估网络流量的健康程度.

(下转第 65 页)

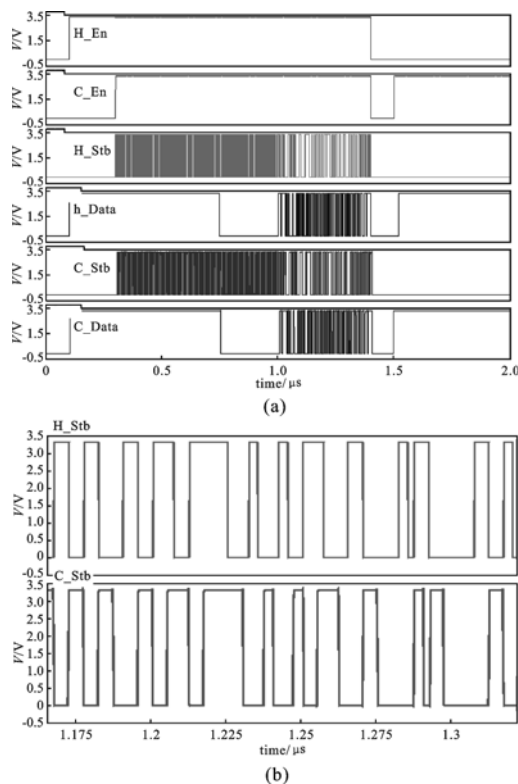


图8 系统仿真验证结果

低了驱动器对温度和工艺的灵敏度. 高速接收器采用单个PMOS差分对作为输入级,电路结构简单,可节省芯片面积. 低功耗接收器采用DDC结构,差动接收输入信号,提高了低功耗接收器抑制共模噪声的能力. 经仿真验证,MDDI主/客端的物理层传输电路传输速率高达400 Mb/s,在低功耗模式下,主客端均可检测启动序列,唤醒高速链接,主客端的

静态功耗均为0.75 mW,完全满足MDDI协议的要求,该电路已成功应用于AM-OLED驱动芯片中.

参考文献:

- [1] 熊文彬. 基于FPGA的OLED显示系统[D]. 成都: 电子科技大学, 2011.
- [2] VESA Mobile display digital interface standard[S]. VESA, 2008.
- [3] Rafal Kleczel. The design of low power 11.6 mW high speed 1.8 Gb/s stand-alone LVDS driver in 0.18 μm CMOS[C]//MIXDES 2010, 17th International Conference on Mixed Design of Integrated Circuits and Systems, Wroclaw, Poland, 2010.
- [4] Hari Shanker Gupta, Parmar R M, Dave R K. High speed LVDS driver for SERDES[C]//International Conference on Emerging Trends in Electronic and Photonic Devices & Systems, Varanasi, 2009.
- [5] Carlos Z R, Teresa S G, Bernable L B. A 0.35 μm sub-ns wake-up time on-off switchable LVDS driver-receiver chip I/O pad pair for rate-dependent power saving in ARE bit-serial links[J]. IEEE Transactions on Biomedical Circuits and Systems, 2011, 6(5): 486-497.
- [6] Eung-Ju Kim, Ho-Young Park. Unified dual mode physical layer for mobile CMOS image sensor interface[J]. IEEE Transactions on Consumer Electronics, 2010, 56(3): 1196-1203.

作者简介:

夏佳佳 女, (1988—), 硕士研究生. 研究方向为模拟与混合信号VLSI设计.

(上接第60页)

参考文献:

- [1] Du X, Yang Y J, Kang X W. Research of applying information entropy and clustering technique on network traffic analysis[C]// Proceedings of the 2008 International Conference on Computational Intelligence and Security, Suzhou: IEEE, 2008: 482-486.
- [2] Joarder R A, Parveen S, Sarwar H, et al. Analysis of real-time multimedia traffic in the context of self-similarity[C]// Proceedings of the 5th International Conference on Electrical and Computer Engineering, Dhaka: IEEE, 2008: 618-623.
- [3] Dainotti A, Pescapé A, Ventre G. Worm traffic analysis and characterization[C]// Proceedings of IEEE ICC, Scotland, Glasgow, 2007: 1435-1442.
- [4] Costa M, Goldberger A L, Peng C K. Multiscale entropy analysis of biological signals[J]. Physical Review E, 2005(71, 021906): 1-18.

- [5] Lakhina A, Crovella M, DiO C. Characterization of network-wide anomalies in traffic flows[C]// Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, New York, 2004: 201-206.
- [6] Soule A, Salamatian K, Taft N. Combining filtering and statistical methods for anomaly detection[C]// Proceedings of the USENIX IMC, Philadelphia: ACM, 2005: 331-344.
- [7] Abilene data [DB/OL]. [2010-05-20]. <http://www.cs.utexas.edu/~yzhang/research/AbileneTM/X01.gz>

作者简介:

颜若愚 男, (1974—), 博士, 讲师. 研究方向为网络安全.