



国际信息工程先进技术译丛



云应用中的服务质量

Service Quality Of Cloud-Based Applications

[美] 埃里克·鲍尔 (Eric Bauer) 著
兰迪·亚当斯 (Randee Adams)
谭励 杨明华 译

Service Quality of Cloud-Based Applications

Eric Bauer
Randee Adams



WILEY

- ◎在云应用程序设计上的架构和工程技术方面给出专业化建议
- ◎帮助架构师、开发人员和测试人员为客户和终端用户开发出高质量的应用



WILEY



机械工业出版社
CHINA MACHINE PRESS

国际信息工程先进技术译丛

云应用中的服务质量

[美] 埃里克·鲍尔 (Eric Bauer) 著
兰迪·亚当斯 (Randee Adams)
谭 励 杨明华 译



机械工业出版社

本书通过“配置”、“分析”和“建议”三个部分，先介绍了关于应用程序服务质量、云模型以及虚拟化架构缺陷的基础内容，然后系统地分析了应用程序服务由于云架构缺陷受到的影响，继而为云计算应用以及尚在开发过程中的应用提供了技术和策略方面的建议，最大化其能够提供优质服务的能力，使通过云计算架构交付给用户的软件应用和服务，具有与在传统本地硬件配置上运行时相同级别的服务质量、可靠性和可用性。

本书能够帮助应用的架构师、开发人员和测试人员为客户和终端用户开发出符合期望，满足要求的高质量应用。适合从事云计算、云应用设计以及软件工程行业的人士阅读，也适合作为相关专业的师生的参考书。

Copyright© 2014 John Wiley & Sons, Ltd.

All Right Reserved. This translation published under license. Authorized translation from English language edition, entitled < Service Quality of Cloud-Based Applications >, ISBN: 978-1-118-76329-2, by Eric Bauer and Randee Adams, Published by John Wiley & Sons. No part of this book may be reproduced in any form without the written permission of the original copyrights holder.

本书中文简体字版由机械工业出版社出版, 未经出版者书面允许, 本书的任何部分不得以任何方式复制或抄袭。版权所有, 翻印必究。

北京市版权局著作权合同登记 图字: 01-2014-6480 号。

图书在版编目 (CIP) 数据

云应用中的服务质量/(美)鲍尔 (Bauer, E.), (美)亚当斯 (Adams, R.) 著; 谭励, 杨明华译. —北京: 机械工业出版社, 2016. 1

(国际信息工程先进技术译丛)

书名原文: Service Quality Of Cloud-Based Applications

ISBN 978-7-111-52352-9

I. ①云… II. ①鲍…②亚…③谭…④杨… III. ①计算机网络
IV. ①TP393

中国版本图书馆 CIP 数据核字 (2015) 第 301096 号

机械工业出版社 (北京市百万庄大街 22 号 邮政编码 100037)

策划编辑: 吕 潇 责任编辑: 吕 潇

责任校对: 闫玥红 责任印制: 李 洋

三河市国英印务有限公司印刷

2016 年 1 月第 1 版第 1 次印刷

169mm×239mm·16.5 印张·338 千字

0001—3000 册

标准书号: ISBN 978-7-111-52352-9

定价: 78.00 元

凡购本书, 如有缺页、倒页、脱页, 由本社发行部调换

电话服务

网络服务

服务咨询热线: 010-88361066

机工官网: www.cmpbook.com

读者购书热线: 010-68326294

机工官博: weibo.com/cmp1952

010-88379203

金书网: www.golden-book.com

封面无防伪标均为盗版

教育服务网: www.cmpedu.com

关于作者

Eric Bauer 是阿尔卡特-朗讯的 IP 平台 CTO 的可靠性工程经理，他曾在阿尔卡特-朗讯的平台、应用以及解决方案的可靠性方面工作超过十年。在从事可靠性工程领域之前，Bauer 花了二十年时间设计和开发嵌入式固件、网络操作系统、IP PBX、互联网平台以及光传输系统。Bauer 获得了十多项美国专利，撰写了《Reliability and Availability of Cloud Computing (云计算实战：可靠性与可用性设计)》《Beyond Redundancy: How Geographic Redundancy Can Improve Service Availability and Reliability of Computer-Based Systems (超越冗余：地理冗余如何才能提高计算机系统的可用性和可靠性)》《Design for Reliability: Information and Computer-Based Systems (可靠性设计：信息和计算机系统)》《Practical System Reliability (系统可靠性实用技术)》(均由 Wiley-IEEE 出版社出版)等著作，并有多篇论文在《Bell Labs Technical Journal (贝尔实验室技术期刊)》发表。Bauer 拥有康奈尔大学电子工程学士学位和普渡大学电气工程硕士学位，他住在新泽西州弗里霍尔德。

Randee Adams 是阿尔卡特-朗讯的 IP 平台 CTO 的技术顾问，她花了近十年时间专注于产品的可靠性设计，曾多次在各种内部可靠性论坛上发言。Adams 撰写了《Beyond Redundancy: How Geographic Redundancy Can Improve Service Availability and Reliability of Computer-Based Systems (超越冗余：地理冗余如何才能提高计算机系统的可用性和可靠性)》和《Reliability and Availability of Cloud Computing (云计算实战：可靠性和可用性设计)》等著作。她最初作为 5ESS 交换机的程序员，于 1979 年加入贝尔实验室。Adams 在整个公司的多个项目（如软件开发、故障单管理、负载管理研究、软件交付、系统工程、软件架构、软件设计、开发工具和联合风险设置）和多个功能领域（如数据库管理、公共信道信令、操作实施、指导和管理、可靠性和安全性）工作过。Adams 拥有亚利桑那大学的学士学位以及伊利诺伊理工学院的计算机科学硕士学位，她住在伊利诺伊州内珀维尔。

译者序

本书的作者为我们规划了部署在云上的应用程序的美好愿景，那就是这些在云架构上的应用和服务，应该能够与部署在传统、本地的硬件上一样，具有良好的服务质量、可靠性和可用性。云计算架构在具有优势的同时也带来了一系列在虚拟化计算、内存、存储和网络资源损耗方面的风险和缺陷，应用程序开发人员和运营商应当尽可能避免云计算架构的缺陷，才能保证应用和服务交付给最终用户时不会受到很大影响。

本书介绍了云模型和基于云的应用程序服务质量，分析了可能影响交付给最终用户的应用程序服务质量的虚拟化架构缺陷，并探讨了改进云服务质量的各种可能。同时，本书还推荐了一些关于架构、策略和相关技术方面的建议，帮助读者在云计算应用程序开发和部署过程中，实现服务质量方面的优化。

本书由三个部分组成：配置、分析与建议。第一部分主要是一些概念，介绍了什么是服务质量，什么是云模型，什么是虚拟化架构的缺陷等。第二部分则分析了虚拟化架构的缺陷是如何影响应用程序服务质量的，分析涉及冗余、负载均衡、版本管理、容量管理等多个方面。第三部分则是作者的建议，如何能够使云应用满足服务质量方面的需求。作者在服务质量方面的从业经验丰富，书中经常能够用具体生动的例子对概念进行解释，不仅如此，书中还提供了大量的交叉引用，方便读者能够前后查找一些概念或者有选择性的对本书进行阅读。

希望读者能在本书中找到对自己有用的东西，为开发和部署一个高服务质量的云应用做些准备。本书由谭励和杨明华合译，水平有限，虽然不是第一次翻译外文文献，但本书绝对是用时最长的，特别是在处理和核对书中交叉引用的内容上，为了保证前后文的一致性，的确花费了大量的时间。尽管如此，书中仍难免有各种翻译不当的地方，还请各位同仁指正。

译者
2015 年底

目 录

关于作者

译者序

第1章 概述	1
1.1 入门	1
1.2 目标读者	2
1.3 本书组织结构	2

I 配 置

第2章 应用程序服务质量	6
2.1 简单应用程序模型	6
2.2 服务边界	7
2.3 质量和性能的关键指标	8
2.4 关键应用特征	11
2.4.1 服务急迫性	12
2.4.2 应用程序交互性	12
2.4.3 网络传输缺陷的耐受性	13
2.5 应用程序服务质量指标	14
2.5.1 服务可用性	14
2.5.2 服务延迟	15
2.5.3 服务可靠性	20
2.5.4 服务可访问性	20
2.5.5 服务可维持性	20
2.5.6 服务吞吐量	21
2.5.7 服务时间戳精度	21
2.5.8 特定应用程序的服务质量度量	21
2.6 技术服务与支持服务	22
2.6.1 技术服务质量	22
2.6.2 支持服务质量	22
2.7 安全事项	23

第3章 云模型	24
---------------	----

3.1	云计算中的角色	24
3.2	云服务模型	25
3.3	云的基本特征	25
3.3.1	按需自助服务	25
3.3.2	广泛的网络访问	26
3.3.3	资源池	26
3.3.4	快速弹性	26
3.3.5	度量服务	27
3.4	简化云架构	27
3.4.1	应用软件	27
3.4.2	虚拟机服务器	28
3.4.3	虚拟机服务器控制器	29
3.4.4	云操作支持系统	29
3.4.5	云技术组件“即服务”	29
3.5	弹性度量	30
3.5.1	密度	30
3.5.2	配置间隔	31
3.5.3	释放间隔	32
3.5.4	向内和向外扩展	32
3.5.5	向上和向下扩展	34
3.5.6	敏捷性	34
3.5.7	转换速率和线性度	35
3.5.8	弹性加速	36
3.6	空间和区域	37
3.7	云意识	38
第4章	虚拟化架构缺陷	40
4.1	服务延迟、虚拟化和云	41
4.1.1	虚拟化和云导致的延迟变化	41
4.1.2	虚拟化开销	42
4.1.3	增加架构性能的可变性	43
4.2	虚拟机故障	43
4.3	无法交付的虚拟机配置容量	44
4.4	交付退化的虚拟机容量	46
4.5	尾部延迟	48
4.6	时钟事件抖动	49
4.7	时钟漂移	50
4.8	失败或缓慢的虚拟机实例分配和启动	51

4.9 虚拟化架构缺陷展望	51
---------------------	----

II 分 析

第5章 应用程序冗余和云计算	54
5.1 故障、可用性和简单架构	54
5.2 通过虚拟化改进软件修复时间	56
5.3 通过虚拟化改进架构修复时间	57
5.3.1 理解硬件修复	57
5.3.2 虚拟机修复即服务	58
5.3.3 讨论	60
5.4 冗余和可恢复性	60
5.4.1 通过虚拟化改进恢复时间	64
5.5 顺序冗余和并发冗余	65
5.5.1 混合并发策略	68
5.6 虚拟化缺陷对应用服务的影响	69
5.6.1 简单架构的服务影响	69
5.6.2 顺序冗余架构的服务影响	69
5.6.3 并发冗余架构的服务影响	71
5.6.4 混合并发架构的服务影响	72
5.7 数据冗余	74
5.7.1 数据存储策略	74
5.7.2 数据一致性策略	75
5.7.3 数据架构注意事项	76
5.8 讨论	76
5.8.1 服务质量的影响	76
5.8.2 并发控制	77
5.8.3 资源使用	77
5.8.4 简易性	78
5.8.5 其他注意事项	78
第6章 负载分配与均衡	79
6.1 负载分配机制	79
6.2 负载分配策略	80
6.3 代理负载均衡器	81
6.4 非代理负载分配	82
6.5 负载分配的层次结构	83
6.6 基于云的负载均衡所面临的挑战	83
6.7 负载均衡在支持冗余方面的作用	84

6.8 负载均衡与可用区域	84
6.9 工作负载服务度量	85
6.10 操作注意事项	86
6.10.1 负载均衡与弹性	86
6.10.2 负载均衡与过载	86
6.10.3 负载均衡与发布管理	87
6.11 负载均衡与应用程序服务质量	87
6.11.1 服务可用性	87
6.11.2 服务延迟	88
6.11.3 服务可靠性	88
6.11.4 服务可访问性	88
6.11.5 服务可维持性	89
6.11.6 服务吞吐量	89
6.11.7 服务时间戳精度	89
第7章 故障容器	90
7.1 故障容器	90
7.1.1 故障级联	90
7.1.2 故障容器与恢复	91
7.1.3 故障容器与虚拟化	92
7.2 故障点	93
7.2.1 单点故障	93
7.2.2 单点故障与虚拟化	95
7.2.3 关联性和反关联性考虑	97
7.2.4 在云计算中确保无 SPOF	97
7.2.5 无 SPOF 和应用程序数据	98
7.3 极端共存解决方案	99
7.3.1 极端共存解决方案的风险	100
7.4 多租户与解决方案容器	101
第8章 容量管理	102
8.1 工作负载变化	102
8.2 传统容量管理	103
8.3 传统过载控制	104
8.4 容量管理与虚拟化	105
8.5 云容量管理	106
8.6 弹性存储注意事项	109
8.7 弹性和过载	109

8.8 操作注意事项	110
8.9 负载拉锯	112
8.10 一般弹性风险	112
8.11 弹性故障场景	113
8.11.1 弹性增长故障场景	113
8.11.2 弹性容量逆增长故障场景	115
第9章 发布管理	117
9.1 相关术语	117
9.2 传统的软件升级策略	117
9.2.1 软件升级需求	118
9.2.2 维护窗口	119
9.2.3 应用升级的客户端注意事项	120
9.2.4 传统的离线软件升级	120
9.2.5 传统的在线软件升级	121
9.2.6 讨论	122
9.3 支持云的软件升级策略	123
9.3.1 I型云支持升级策略：街区聚会	124
9.3.2 II型云支持升级策略：每车一司机	125
9.3.3 讨论	126
9.4 数据管理	127
9.5 软件升级中的服务编排角色	128
9.5.1 解决方案级软件升级	129
9.6 结论	129
第10章 端到端考虑因素	130
10.1 端到端服务环境	130
10.2 三层端到端服务模型	135
10.2.1 通过三层模型估算服务缺陷	136
10.2.2 端到端服务可用性	137
10.2.3 端到端服务延迟	138
10.2.4 端到端服务可靠性	139
10.2.5 端到端服务可访问性	140
10.2.6 端到端服务可维持性	140
10.2.7 端到端服务吞吐量	141
10.2.8 端到端服务时间戳精度	141
10.2.9 现实检查	141
10.3 分布式和集中式的云数据中心	142

10.3.1	集中式云数据中心	142
10.3.2	分布式云数据中心	142
10.3.3	服务可用性考虑	143
10.3.4	服务延迟考虑	145
10.3.5	服务可靠性考虑	145
10.3.6	服务可访问性考虑	146
10.3.7	服务可维持性考虑	146
10.3.8	资源分配考虑	146
10.4	多层解决方案架构	147
10.5	灾难恢复与地理冗余	148
10.5.1	灾难恢复目标	148
10.5.2	地理冗余架构	149
10.5.3	服务质量考虑	149
10.5.4	恢复点考虑	150
10.5.5	地理冗余和可用区域减轻灾难的影响	151

III 建 议

第 11 章	服务质量问责	154
11.1	传统的问责	154
11.2	云服务交付路径	155
11.3	云问责	157
11.4	问责案例研究	159
11.4.1	问责和技术组件	160
11.4.2	问责和弹性	162
11.5	服务质量差距模型	163
11.5.1	应用程序面向资源服务差距分析	164
11.5.2	应用程序面向用户服务差距分析	166
11.6	服务水平协议	168
第 12 章	服务可用性度量	170
12.1	服务度量概述	170
12.2	传统服务可用性度量	171
12.3	服务可用性度量演化	172
12.3.1	应用演化分析	173
12.3.2	技术组件	178
12.3.3	存储即服务的使用	179
12.4	硬件可靠性度量演化	180
12.4.1	虚拟机故障生命周期	181

12.5	弹性服务可用性度量演化	182
12.6	发布管理服务可用性度量演化	183
12.7	服务度量展望	185
第 13 章	应用程序服务质量需求	186
13.1	服务可用性需求	186
13.2	服务延迟需求	189
13.3	服务可靠性需求	189
13.4	服务可访问性需求	190
13.5	服务可持续性需求	190
13.6	服务吞吐量需求	191
13.7	时间戳精度需求	191
13.8	弹性需求	191
13.9	发布管理需求	192
13.10	灾难恢复需求	192
第 14 章	虚拟化架构度量与管理	194
14.1	架构服务质量度量的业务环境	194
14.2	云消费者的度量选择	195
14.3	缺陷度量策略	197
14.3.1	虚拟机故障度量	197
14.3.2	无法交付的虚拟机配置容量度量	198
14.3.3	交付退化的虚拟机容量度量	198
14.3.4	尾部延迟度量	198
14.3.5	时钟事件抖动度量	199
14.3.6	时钟漂移度量	199
14.3.7	失败或缓慢的虚拟机实例分配和启动度量	199
14.3.8	度量总结	200
14.4	管理虚拟化架构缺陷	201
14.4.1	最小化应用程序对架构缺陷的敏感度	201
14.4.2	虚拟机级拥塞检测与控制	201
14.4.3	分配更多虚拟资源容量	202
14.4.4	终止性能欠佳的虚拟机实例	202
14.4.5	接受性能退化	202
14.4.6	积极主动的供应商管理	202
14.4.7	重新设定最终用户服务质量期望	202
14.4.8	SLA 注意事项	203
14.4.9	更换云服务提供商	203

第 15 章 基于云的应用程序分析	204
15.1 可靠性框图和参照分析	204
15.2 IaaS 缺陷影响分析	205
15.3 PaaS 故障影响分析	207
15.4 工作负载分配分析	208
15.4.1 服务质量分析	208
15.4.2 过载控制分析	209
15.5 反关联性分析	209
15.6 弹性分析	210
15.6.1 服务容量增长场景	211
15.6.2 服务容量增长操作分析	211
15.6.3 服务容量逆增长操作分析	212
15.6.4 存储容量增长场景	212
15.6.5 在线存储容量增长操作分析	213
15.6.6 在线存储容量逆增长操作分析	213
15.7 发布管理影响效应分析	213
15.7.1 服务可用性影响	213
15.7.2 服务可靠性影响	214
15.7.3 服务可访问性影响	214
15.7.4 服务可维持性影响	214
15.7.5 服务吞吐量影响	214
15.8 恢复点目标分析	214
15.9 恢复时间目标分析	216
第 16 章 测试注意事项	218
16.1 测试环境	218
16.2 测试策略	218
16.2.1 云测试平台	219
16.2.2 用于测试的容量	219
16.2.3 统计置信度	220
16.2.4 服务中断时间	220
16.3 模拟架构缺陷	221
16.4 测试计划	222
16.4.1 服务可靠性和延迟测试	222
16.4.2 架构缺陷测试	223
16.4.3 健壮性测试	223
16.4.4 持久性/稳定性测试	225

16.4.5 应用程序弹性测试	227
16.4.6 升级测试	228
16.4.7 灾难恢复测试	228
16.4.8 极限共存测试	228
16.4.9 PaaS 技术组件测试	229
16.4.10 自动回归测试	229
16.4.11 构造发布测试	229
第 17 章 关键点连接与总结	230
17.1 应用程序服务质量所面临的挑战	230
17.2 冗余和健壮性	231
17.3 可伸缩性设计	234
17.4 可扩展性设计	234
17.5 故障设计	235
17.6 规划注意事项	236
17.7 传统应用的演化	237
17.7.1 阶段 0: 传统应用	239
17.7.2 阶段 I: 虚拟化架构上的高服务质量	239
17.7.3 阶段 II: 手动应用弹性	240
17.7.4 阶段 III: 自动发布管理	240
17.7.5 阶段 IV: 自动应用弹性	240
17.7.6 阶段 V: 虚拟机迁移	241
17.8 结束语	241
缩略语	242
参考文献	245

第 1 章 概 述

用户希望部署在云计算架构上的应用和服务能够与部署在传统、本地的硬件上一样，具有相似的服务质量、可靠性、可用性和延迟。云计算架构引入了一系列由于虚拟化计算、内存、存储和由“架构即服务（Infrastructure-as-a-Service, IaaS）”供应商带给托管的应用程序实例等网络资源带来的服务缺陷风险，因此，应用程序开发人员和云消费者应当尽可能避免这些缺陷，以确保应用程序服务交付给最终用户时不会受到很大影响。本书分析了可能影响应用程序服务交付给最终用户的云架构问题，以及改进云服务质量的各種可能。同时，本书还推荐了一些架构、策略和相关技术，能够使得部署在云上的应用程序为终端用户提供更好的服务。

1.1 入门

基于云的应用软件在一系列虚拟机实例中执行，每一个独立的虚拟机实例依靠云架构所提供的虚拟计算、内存、存储和网络来进行服务交付。如图 1.1 所示，应用程序通过虚线边界向终端用户提供“面向用户的服务（customer facing service）”，“IaaS”供应商通过图中的虚线边界，即“面向资源服务（resource facing service）”提供虚拟化资源。对于终端用户而言，应用程序的服务质量可以看做是一个由应用程序架构和软件质量构成的函数，而由 IaaS 通过面向资源服务边界提供的虚拟架构的服务质量，以及将终端用户连接至应用程序实例的接入服务和广域网服务质量也是如此。本书考虑了为云应用程序所提供的虚拟化资源存在的各种缺陷，并讨论如何将终端用户体验的用户服务质量最优化。如果忽略终端用户设备的服务缺陷，在接入和广域网中，用户可以勉强感受到应用程序服务质量的差异，从而区分一个特定的应用程序是部署在云架构上的还是部署在传统的硬件设备上的。

应用软件部署在本地或云端的关键技术差异在于，本地部署应用程序的用户操作系统能够直接访问物理计算、内存、存储和网络资源，而云端部署则在用户操作系统和物理硬件之间插入了一个管理程序层或者虚拟机管理软件。这个管理程序层或虚拟机管理软件能够实现复杂的资源共享，技术参数和操作策略。然而，管理程序层或虚拟机管理软件并不能向用户操作系统和应用软件提供合适的硬件仿真，这使得提供给最终用户的应用程序服务质量会受到一定影响。如图 1.1 所示，应用程序部署在一个独立的数据中心，而现实中应用程序往往需要部署在多个数据中心，通过缩短消息抵达最终用户的延迟，支持连续性业务和灾难恢复以及其他商业措

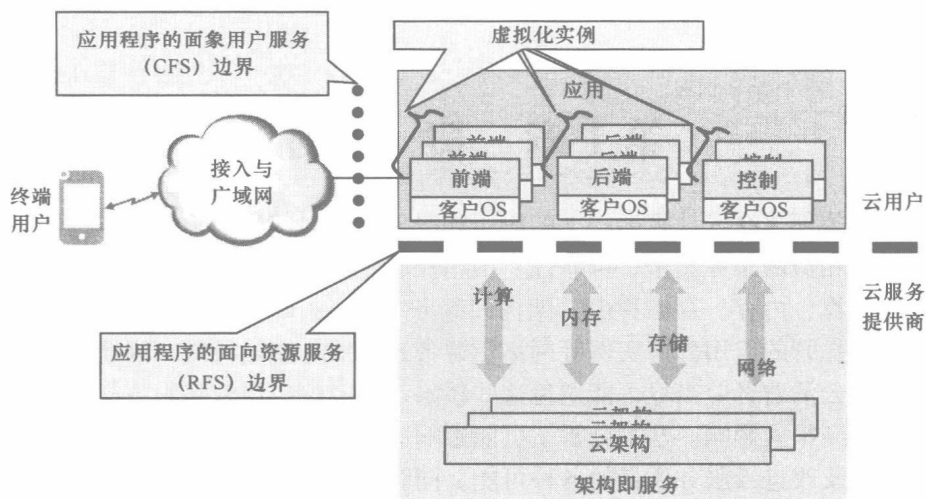


图 1.1 基于云的应用示例

施，才能保证用户的服务质量。本书也会涉及部署在多个数据中心的应用程序服务质量问题。

本书提供了当应用程序软件部署在云架构上时，为保证交付给最终用户好的应用程序服务质量所应采用的应用程序架构、配置、验证和操作策略。本书所采用的保障应用程序服务质量的方法，来自于终端用户视角，同时还参考了行业标准和来自 NIST、TM 论坛、QuEST 论坛、ODCA、ISO、ITIL 等联盟的推荐。

1.2 目标读者

本书为应用程序架构师、开发人员和测试人员提供了设计和工程应用的指导，能够满足客户和最终用户在服务可靠性、可用性、质量和延迟方面的期望。产品经理、开发经理和项目经理也将从本书中获得关于服务质量风险方面的深入理解，风险必须尽可能减小才能确保一个应用程序部署到云架构时，能够一如既往地满足或超过客户在用户服务质量方面的预期。

1.3 本书组织结构

本书由三个部分组成：配置、分析与建议。

第 I 部分：配置，将基于云的应用程序服务质量配置做出划分：

- “应用程序服务质量（第 2 章）”。本章定义了书中对于应用程序服务质量的度量标准，包括：服务可用性、服务延迟、服务可靠性、服务可访问性、服务可维持性、服务吞吐量以及服务时间戳精度。