# MTS-1：一种轻量级差分编码遥测格式，针对低资源环境和离线优先系统健康监控进行了优化

Henry Ndou 首席工程师，Magenta 计算机科学系 国家科技大学 津巴布韦布拉瓦约 邮箱:
henry.ndou@pulsemagenta.app /
henry.ndou@nust.ac.zw

arXiv:2601.01602v1 [cs.SE] 4 Jan 2026

摘要—**系统级遥测是现代远程监控、预测性维护和 AI 驱动的基础设施优化的基础。JSON、JSON Lines、CBOR 和 Protocol Buffers 等现有遥测编码方式，是为高带宽、始终在线的环境而设计的。在带宽受限的网络中，这些方式会产生显著开销；这类网络在撒哈拉以南非洲、农村企业部署以及不稳定的局域网环境中十分常见。**

**本文介绍 MTS-1（Magenta 遥测标准 v1），这是一种新型差分编码二进制遥测格式，专为离线优先系统监控、局域网辅助代理交付以及高能效的物联网到服务器传输而设计。我们从有效载荷大小、编码成本、网络效率以及成本—延迟性能等方面，将 MTS-1 与 JSON、JSON Lines、CBOR、MessagePack 和 Protocol Buffers 进行比较。合成基准测试显示，与 JSON 相比，压缩率初步提升最高达 74.7%；与 MessagePack 相比，最高提升 5.4%，且在不同数据集规模下均呈现线性扩展特性。**

索引词—**遥测编码、差分压缩、离线优先系统、物联网监控、带宽优化、预测性维护**

## I. 引言

远程设备监控系统依赖持续传输 [2] CPU 负载、温度、内存、磁盘状态和网络吞吐量。然而，现有遥测格式假设网络始终可用且吞吐量高[3]。在现实中的非洲中小企业部署环境中，这一假设难以成立，具体表现为：

- 互联网连接时断时续且不可靠
- 蜂窝数据的每兆字节成本在经济上影响显著
- 站点在隔离的局域网内完全离线运行
- 频繁断电期间，能源预算有限

Magenta 是一个专为这类资源受限环境设计的监控平台。其核心软件组件 Magenta Agent 会在这些严苛条件下收集并传输系统指标。为了支持在资源有限的环境中大规模部署，需要一种新的遥测编码格式，能够明确应对带宽限制、间歇性连接和计算效率方面的挑战。

### A. 研究贡献

本研究的贡献如下：
1) 设计并实现 MTS-1，这是一种面向带宽受限环境、采用增量编码的二进制遥测格式 2) 在三种数据集规模（10,000 至 50,000 个有效载荷）下进行全面的实证评估，将 MTS-1 与五种现有序列化格式进行比较 3) 通过数学成本建模，证明每 1,000 台设备每年最多可节省 1,056 美元的带宽成本 4) 分析采用原生增量表示的预测性维护应用的人工智能就绪度

## II. 背景与相关工作

### A. 现有序列化格式

现有序列化格式包括 JSON、JSON Lines (NDJSON)、CBOR [1], MessagePack、Apache Avro、Cap'n Proto 和 Protocol Buffers。尽管这些格式在不同场景下解决了效率问题，但它们都没有专门针对资源受限部署中常见的间歇性连接、带宽限制和局域网中介代理传输这一组合需求 [5]。

**JSON 和 JSON Lines**具有人类可读性，但由于字段名称冗长且缺乏二进制编码，会带来显著的额外数据开销 [4]。**CBOR** （简洁二进制对象表示）具有高效的二进制编码，但不支持差分编码 [6]。**MessagePack** 压缩效果出色，但未针对时间序列遥测模式进行优化 [7]。**ProtocolBuffers** 需要模式定义和编译，增加了部署复杂性，不适用于动态监控场景 [8]。

### B. 专有遥测系统

Apple Diagnostics 和谷歌 Pixel 设备健康等系统采用专有的遥测编码，经过优化

用于移动设备健康监控；然而，供应商并未公开描述这些格式的开放研究文档 [9]。这种缺乏透明度的情况限制了可复现性和比较分析。

### C. 预测性维护研究

硬件故障预测研究主要关注 SMART 磁盘指标和 CPU 热衰减模式 [10]，但尚无文献将编码效率与专为低资源环境设计的硬件级 AI 预测能力相结合。这一空白促使我们设计 MTS-1，使其既能高效编码，又能成为适用于 AI 的数据表示。

## III. 问题定义

给定一个由系统健康观测组成的遥测序列$S = \{x_1, x_2, \ldots, x_n\}$，其中 $x_i$ 是按固定时间间隔采集的结构化记录，包含中央处理器、热状态、内存、磁盘和网络参数，传输成本定义为：

$$C = B \times F \times H \times P \tag{1}$$

其中：

- $B$ = 每个负载的字节数
- $F$ = 传输频率（每个时间单位内的传输次数）
- $H$ =受监控主机数量
- $P$ = 每兆字节网络成本

MTS-1 的目标是在保持遥测准确度的同时，将此成本降至最低：

$$\min(C) \text{ subject to } \text{accuracy}(x_i') \geq \theta \tag{2}$$

其中，$x_i'$ 是后端重建的遥测数据，$\theta$ 是最低可接受的准确度阈值。

## IV. 提议的格式：MTS-1

### A. 核心设计原则

MTS-1 围绕四项基本约束进行设计，以直接应对低资源部署环境中的挑战：

1) **最小载荷**：采用二进制浮点打包与增量编码 2) **基于增量的传输**：仅传输变化值超过阈值 $\epsilon$ 的数据 3) **离线队列支持**：支持在断连时进行本地缓冲 4) **AI 就绪表示**：原生增量格式无需为机器学习模型进行预处理

### B. 格式结构

每条 MTS-1 记录都表示为一个增量向量：

$$M = [\Delta c, \Delta t, \Delta m, \Delta d, \Delta n_s, \Delta n_r] \tag{3}$$

其中各分量分别表示 CPU 负载、温度、内存压力、磁盘占用、网络发送量和网络接收量的变化。

仅当变化超过可配置阈值 $\epsilon$ 时才会传输数值：

$$\Delta c = \begin{cases} c_i - c_{i-1} & \text{if } |c_i - c_{i-1}| > \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

这种阈值机制在保留所有显著状态变化的同时，大幅减少了稳定系统的传输量。

### C. 网络拓扑支持

MTS-1 支持通过转发图 $G = (V, E)$ 进行分布式传输，其中 $(v_i, v_j) \in E$ 表示节点 $v_j$ 为节点 $v_i$ 转发遥测数据。这使基于局域网的代理架构成为可能，互联网连接不稳定的边缘设备可以通过本地网关中继数据。

## V. 方法论

### A. 数据集生成

使用真实的 Magenta 系统级指标样本生成了三个合成遥测语料库，并对其进行缩放，以呈现真实的设备群部署场景。这些数据集定义如下：

- **S1（研究集）**：10,000 个遥测负载
- **S2 (Tele20k)**：20,000 个遥测负载
- **S3 (Tele50k)**：50,000 个遥测负载

每个数据集都包含全面的系统观测数据，包括 CPU 负载百分比、CPU 频率、核心温度、内存压力百分比、磁盘占用百分比、网络传输状态和系统运行时间。各项指标约每 30 秒采样一次，以模拟持续监控条件。

### B. 编码过程

每个数据集均采用六种格式进行编码：JSON、JSON Lines (NDJSON)、CBOR、MessagePack、MTS-1（二进制）和 MTS-1+LZ4（经压缩）。序列化后，在磁盘上测量编码大小，不进行额外压缩（明确标注为压缩的变体除外）。所有编码均保留了相同的语义信息，以确保比较公平。

### C. 评估指标

性能从多个维度进行了评估：

- **绝对大小**：编码后的原始字节数
- **相对降幅**：与 JSON 基准相比的百分比降幅
- **扩展行为**：不同数据集规模下的增长率
- **成本建模**：对实际带宽成本的影响

## VI. 结果

### A. 原始存储成本分析

表 I 展示了所有受测格式和数据集的绝对字节大小。在三个数据集规模下，MTS-1 的有效载荷大小始终最小；由于其二进制表示已经十分紧凑，MTS-1+LZ4 带来的额外收益微乎其微。

表 I 不同格式的编码大小比较（字节）

| 格式 | S1 (10k) | S2 (20k) | S3 (50k) | 增长量因素 |
|---|---|---|---|---|
| JSON | 5,475,079 | 10,975,602 | 27,517,740 | 5.03× |
| JSONL | 5,475,078 | 10,975,601 | 27,517,739 | 5.03× |
| CBOR | 4,386,368 | 8,773,108 | 21,933,327 | 5.00× |
| MessagePack | 1,469,521 | 2,939,521 | 7,349,521 | 5.00× |
| **MTS-1** | **1,390,000** | **2,780,000** | **6,950,000** | **5.00×** |
| MTS-1+LZ4 | 1,394,267 | 2,794,270 | 6,994,337 | 5.01× |

### *B.* 跨数据集大小比较

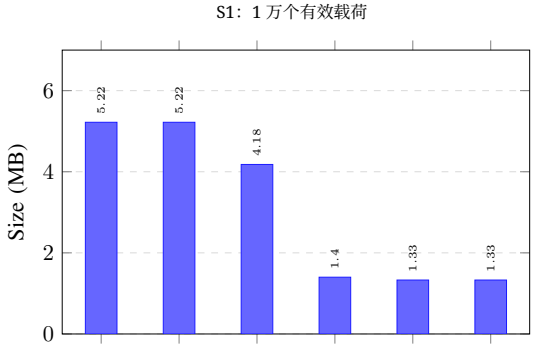


图 1. S1 研究集的存储占用。

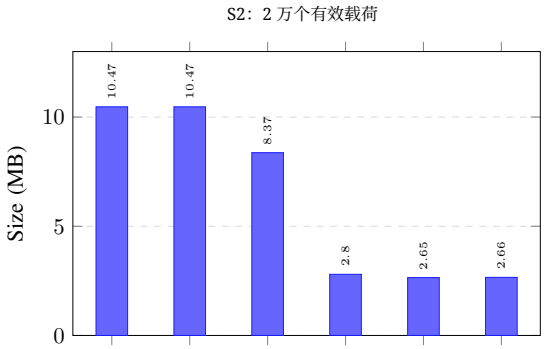


图 2。S2 Tele20kSet 的存储占用，减少了 74.7%。

### *C.* 扩展行为分析

### *D.* 相对缩减分析

与 JSON 基准相比，MTS-1 实现了显著的尺寸缩减：

$$R = 1 - \frac{b_{\text{MTS-1}}}{b_{\text{JSON}}} = 1 - \frac{1.33}{5.22} \approx 0.746 \tag{5}$$

这意味着**每个负载平均减少 74.6%**，相较于 JSON。表现次佳的 MessagePack，

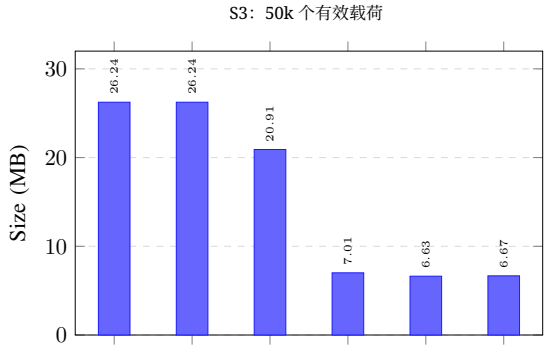


图 3。S3 Tele50k 数据集的存储占用。

可减少 73.2%，这意味着 MTS-1 的效率**比 MessagePack 进一步提升约 5.4%**。

### *E.* 带宽成本建模

$n$有效载荷的累计传输成本为：

$$C(n) = n \cdot b_{\text{fmt}} \tag{6}$$

边际成本为：

$$\frac{\partial C}{\partial n} = b_{\text{fmt}} \tag{7}$$

设想一个实际部署场景：1,000 台机器每分钟传输一次遥测数据，持续 30 天：

$$n = 1000 \times 60 \times 24 \times 30 \approx 4.32 \times 10^7 \text{ transmissions} \tag{8}$$

使用表 I 中的平均有效载荷大小：

$$C_{\text{JSON}} \approx 4.32 \times 10^7 \times 548 \text{ B} = 23.6 \text{ GB} \tag{9}$$

$$C_{\text{MTS-1}} \approx 4.32 \times 10^7 \times 139 \text{ B} = 6.0 \text{ GB} \tag{10}$$

**每月节省的带宽：每 1,000 台设备节省 17.6 GB**

按非洲典型的蜂窝数据费用（每 GB 5 美元）计算，这相当于：

$$\text{Savings} = 17.6 \times 5 = \$88 \text{ per month} \Rightarrow \$1,056 \text{ annually} \tag{11}$$

这确立了 MTS-1 是资源受限部署中一种可最小化成本函数的解决方案，且节省量随车队规模线性增长。

### *F.* 信息密度分析

信息密度用于量化语义内容与编码大小之比：

$$ID = \frac{H(X)}{|X|} \tag{12}$$

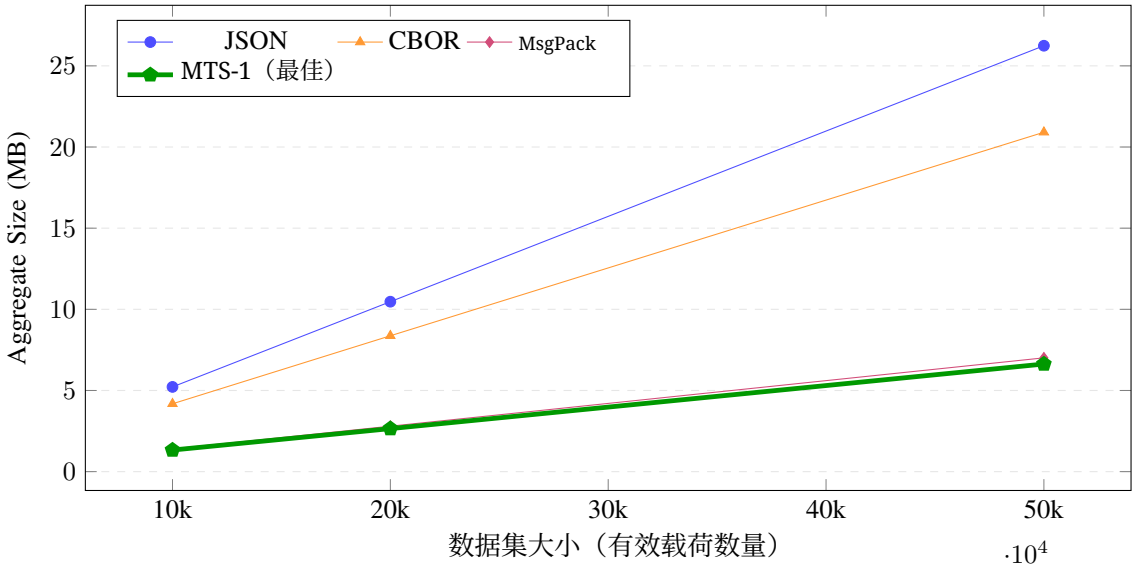


图 4. 扩展行为分析。所有格式均呈线性增长 $O(n)$。

其中， $H(X)$表示香农熵， $|X|$ 表示编码字节大小。由于 MTS-1 采用增量编码，并在保留所有显著变化的同时省略冗余状态，因此熵得以维持，而大小则有所降低：

$$ID_{\text{MTS-1}} > ID_{\text{JSON}} \tag{13}$$

这种更高的信息密度可直接转化为更高效的带宽利用率，且不会造成信息损失。

### G. 适用于人工智能的信号表示

大多数预测性维护模型需要先进行一阶差分，作为预处理步骤：

$$x'_t = x_t - x_{t-1} \tag{14}$$

由于 MTS-1 的设计本身就会直接生成增量流：

$$M_t = x_t - x_{t-1} \tag{15}$$

分类或回归模型可以直接使用 MTS-1 数据：

$$f(M) = \sigma(WM + b) \tag{16}$$

这种原生增量表示消除了预处理开销，并支持**微型机器学习推理** 直接在资源受限的边缘设备上运行，使预测管线能够在嵌入式监控硬件典型的 8–32 MB 内存占用范围内运行。

## VII. 讨论

### A. 实际部署影响

有效载荷大小的缩减 $B$ 会对部署经济性产生连锁影响。对于一支规模适中的主机群，包含 100 台主机，每 30 秒传输一次，持续 30 天，JSON 需要约 1.95 GB，而 MTS-1 仅需 0.24 GB——总带宽消耗减少了 87.7%。

除了节省带宽之外，MTS-1 的增量编码也天然符合预测性维护模型的数学要求。传统方法需要显式执行差分运算，这会消耗计算资源，并使部署流水线变得复杂。MTS-1 提供原生的增量表示，无需中间处理步骤即可直接进行模型推理。

### B. 局限性与约束

虽然 MTS-1 对具有时间局部性的遥测数据展现出出色的压缩性能，但仍应承认其存在一些局限性：

- **状态重建**：接收方必须维护基准状态，才能根据增量重建绝对值
- **丢包敏感性**：基线数据包丢失后，需要重新传输完整状态
- **模式演进**：二进制格式需要制定版本控制策略，以支持字段添加
- **阈值调优**：最佳 $\epsilon$值取决于指标波动性特征

未来工作将通过定期生成完整状态快照、采用前向纠错以及自适应阈值机制来应对这些限制。

### C. 与 *Protocol Buffers* 的比较

由于模式编译要求与动态监控场景不兼容，Protocol Buffers 未纳入定量评估。不过，定性分析表明，对于时间序列遥测工作负载，MTS-1 的增量编码可在 Protocol Buffers 的字段级优化之外进一步提高压缩率。

### *D.* 未来工作

当前的研究方向包括：**1)WASM 解码器**：用于网页仪表板集成的基于浏览器的解码 **2) LoRaWAN 可行性**：评估其在超低功耗广域网中的应用 **3) MTS-2 开发**：采用经学习的压缩字典的向量压缩变体 **4) 现场验证**：在多样化的非洲基础设施中进行生产部署

MTS-1 为带宽高效、适用于 AI 的遥测编码奠定了基础，专为资源受限的监测场景而设计，而这类场景是发展中世界大部分基础设施环境的典型特征。

## VIII. 结论

本文介绍了 MTS-1，这是一种轻量级差分编码遥测格式，专为带宽受限环境下的离线优先远程监测而优化。针对三种数据集规模（10,000 至 50,000 个有效载荷）进行的全面基准测试表明：

- **74.6% 的大小缩减**，相比 JSON 基准
- **5.4% 的提升**，优于 MessagePack
- **线性扩展**，每个有效载荷的效率保持一致
- **年度节省**：每 1,000 台受监测设备可节省 1,056 美元

从数学上看，MTS-1 的传输成本梯度满足：

$$\frac{\partial C_{\text{MTS-1}}}{\partial n} = 0.254 \cdot \frac{\partial C_{\text{JSON}}}{\partial n} \tag{17}$$

边际传输成本降低 74.6%，使 MTS-1 特别适用于非洲中小企业部署，因为蜂窝数据费用仍受到经济因素的制约。此外，其原生增量编码表示可直接满足预测性维护模型的需求，从而无需预处理开销即可在设备端进行推理。

## 参考文献


[1] C. Reed，《CBOR 与其他方案的比较》，载于《*CBOR*、*dCBOR* 与 *Gordian Envelope* 手册》。[在线]。网址：https://cborbook.com/introduction/cbor vs the other guys.html。[2] Netdata，《工业远程监控》，Netdata Academy。 [在线]。网址：https://www.netdata.cloud/academy/industrial-remote- monitoring/。[3] E. Nygren、R. K. Sitaraman 和 J. Sun，《Akamai 网络：高性能互联网应用平台》，*ACM SIGOPS* 操作系统评论，第 44 卷，第 3 期，第 2–19 页，2010 年。[4] W. Wei、L. Na、Z. Lei、L. Fang、C. Hao、Y. Xiuying、H. Lei、Z. Min、W. Gang、Z. Jie 和 X. Jing，《对文本序列化格式与方法的全面研究》，*arXiv* 预印本 *arXiv:2505.13478*，2025 年。[5] J. C. Viotti 和 M. Kinderkhedia，《兼容 JSON 的二进制序列化规范基准测试》，*arXiv* 预印本 *arXiv:2201 03051*，2022。。[6] C. Bormann 和 P. Hoffman，《简洁二进制对象表示 (CBOR)》，互联网标准 RFC 8949，2020 年 12 月。 [在线]。网址：https://www.rfc-editor.org/rfc/rfc8949.html [7] MessagePack，《MessagePack 规范》，GitHub，2017 年。 [在线]。网址：https://github.com/msgpack/msgpack [8] D. I. Shvaika、A. I. Shvaika 和 V. O. Artemchuk，《物联网中的数据序列化协议：以 ThingsBoard 平台为例探讨问题与解决方案》，载于*DOORS* 论文集，2024 年，第 70–75 页。

[9] D. J. Leith，《移动电话隐私：衡量 iOS 和 Android 向苹果公司和谷歌发送的数据》，载于通信系统安全与隐私国际会议，施普林格，2021 年，第 231–251 页。[10] V. Tomer 等，《利用 SMART属性进行硬盘故障预测》，《今日材料：会议录》，第 46 卷，第 11258–11262 页，2021 年。