exterro®
Data Risk Management

FTK SUITE 8.2 SP2

AI SERVER BENCHMARK METRICS GUIDE

AUGUST 2025

Table of Contents

About Exterro

Exterro was founded with the simple vision that applying the concepts of process optimization and data science to how companies manage digital information and respond to litigation would drive more successful outcomes at a lower cost. We remain committed to this vision today. We deliver a fully integrated Data Risk Management platform that enables our clients to address their privacy, regulatory, compliance, digital forensics, and litigation risks more effectively and at lower costs. We provide software solutions that help some of the world's largest organizations, law enforcement and government agencies work smarter, more efficiently, and support the Rule of Law.

Purpose of the Document

The purpose of this document is to provide a comprehensive overview of benchmarking metrics collected across various AI workloads and hardware configurations within the FTK AI Server environment. This guide serves as a reference for assessing the performance and scalability of AI models deployed in FTK environments using different GPU types, FTK build versions, and server specifications.

Key Focus Areas:

- Inference speed benchmarking across GPU types and model sizes.
- Multimedia Summarization (MMS) job performance.
- Text Summarization job performance.
- Hardware and software environment details.

These benchmarks aim to assist engineering, QA, and infrastructure teams in making informed decisions about performance expectations, hardware provisioning, and optimization strategies.

1 Ollama Inference Speed Benchmarking

This section documents the inference performance (measured in tokens/second) of various Large Language Models (LLMs) running on different NVIDIA GPU chips using Ollama.

1.1 Test Environment

- **FTKC Build Version:** 8.2.2.102 SP2
- **AI Server Version:** 2025.07.31.160
- **FTKC Machine Specification:** c5.4xlarge
- **AI Server Machine Specification:** g5.2xlarge

GPU Chip Specifications:

- **GPU Model**: NVIDIA A10G
- **VRAM Memory**: 24 GB

FTKC Server Specifications:

- **Machine Type**: c5.4xlarge
- **Processor**: Intel® Xeon® Platinum 8275CL @ 3.00 GHz
- **Operating System**: Windows Server 2022 Datacenter
- **RAM**: 32 GB
- **CPU Cores**: 16

AI Server Specifications:

- **Machine Type**: g5.2xlarge
- **Processor**: AMD EPYC 7R32 @ 2.80 GHz
- **Operating System**: Windows Server 2022 Datacenter
- **RAM**: 32 GB
- **CPU Cores**: 8

1.2 Inference Speed by GPU Type

1.2.1 NVIDIA T4

Model Name	Parameters	Inference Speed (Tokens/Second)
llama3.2	3B	58
gemma3:4b	4B	42
llama3.2-vision:11b	11B	25
gemma3:12b	12b	17
qwen3:14b	14B	17

1.2.2 NVIDIA V100

Model Name	Parameters	Inference Speed (Tokens/Second)
llama3.2	3B	120
gemma3:4b	4B	84
llama3.2-vision:11b	11B	82
gemma3:12b	12B	48
qwen3:14b	14B	54

1.2.3 NVIDIA P100

Model Name	Parameters	Inference Speed (Tokens/Second)
llama3.2	3B	48
gemma3:4b	4B	38
llama3.2-vision:11b	11B	28
gemma3:12b	12B	17
qwen3:14b	14B	17

1.2.4 NVIDIA A10G

Model Name	Parameters	Inference Speed (Tokens/Second)
llama3.2	3B	120
gemma3:4b	4B	94
llama3.2-vision:11b	11B	73
gemma3:12b	12B	44
qwen3:14b	14B	43

2 Multimedia Summarization (MMS) Metrics

This section benchmarks the AI engine's performance when processing multimedia datasets. Key performance factors include the number of AI servers, object volume, thread count, and job concurrency.

2.1 GPU Metrics

2.1.1 Metrics Captured

- **AI Servers**: Number of AI processing nodes
- **No. of Objects and Document Count**: Volume of input data
- **Threads**: Degree of parallel processing
- **Job Time**: Total processing time (HH:MM:SS)

2.1.2 MMS Job Time – (3) AI Servers

No. of Objects	Document Counts	Threads	Job Time (AI Multimedia) (HH:MM:SS)
10K	10,000	6	00:15:26
		4	00:18:25
		3	00:21:32
50K	50,055	6	01:04:00
		3	01:47:00
100K	100,692	6	02:04:00
		3	03:28:00

2.1.3 MMS Job Time – (6) AI Servers

No. of Objects	Document Counts	Threads	Job Time (AI Multimedia) (HH:MM:SS)
10K	10,000	12	00:15:00
		6	00:14:00
50K	50,055	12	01:04:00
		6	01:05:00
100K	100,692	12	02:08:00
		6	02:14:00

2.2 CPU Metrics

No. of Objects	Job Time (AI Multimedia) (HH:MM:SS)
10K	09:26:03

3 Text Summarization Metrics

This section evaluates the AI engine's performance on document summarization tasks. Metrics vary based on the number of AI servers and input volume.

3.1 GPU Metrics

3.1.1 Metrics Captured

- **AI Servers**: Number of AI processing nodes
- **Objects and Document Count**: Input volume
- **Files Processed**: Total files handled during the job
- **Job Time**: Total processing time (HH:MM:SS)

3.1.2 Job Time – (6) AI Servers

Objects	Document Counts	Files Processed	Job Time (AI Summarization) (HH:MM:SS)
10K	8,440	8,420	01:39:00
50K	58,844	58,812	05:56:00
100K	116,151	115,130	12:52:00

3.1.3 Job Time – (3) AI Servers

Objects	Document Counts	Files Processed	Job Time (AI Summarization) (HH:MM:SS)
10K	8,440	8,420	03:34:00
50K	58,844	58,812	11:52:00
100K	110,355	109,435	22:57:00

3.1.4 Job Time – (1) AI Server

Objects	Document Counts	Files Processed	Job Time (AI Summarization) (HH:MM:SS)
10K	8,440	8,420	08:00:00
50K	58,844	58,812	45:27:00
100K	110,355	109,435	68:04:00

3.2 GPU vs CPU Metrics

A quick metrics-based guide comparing how CPU and GPU handle the same forensic document - **NIST Special Publication 800-86**—in terms of time, output, and depth.

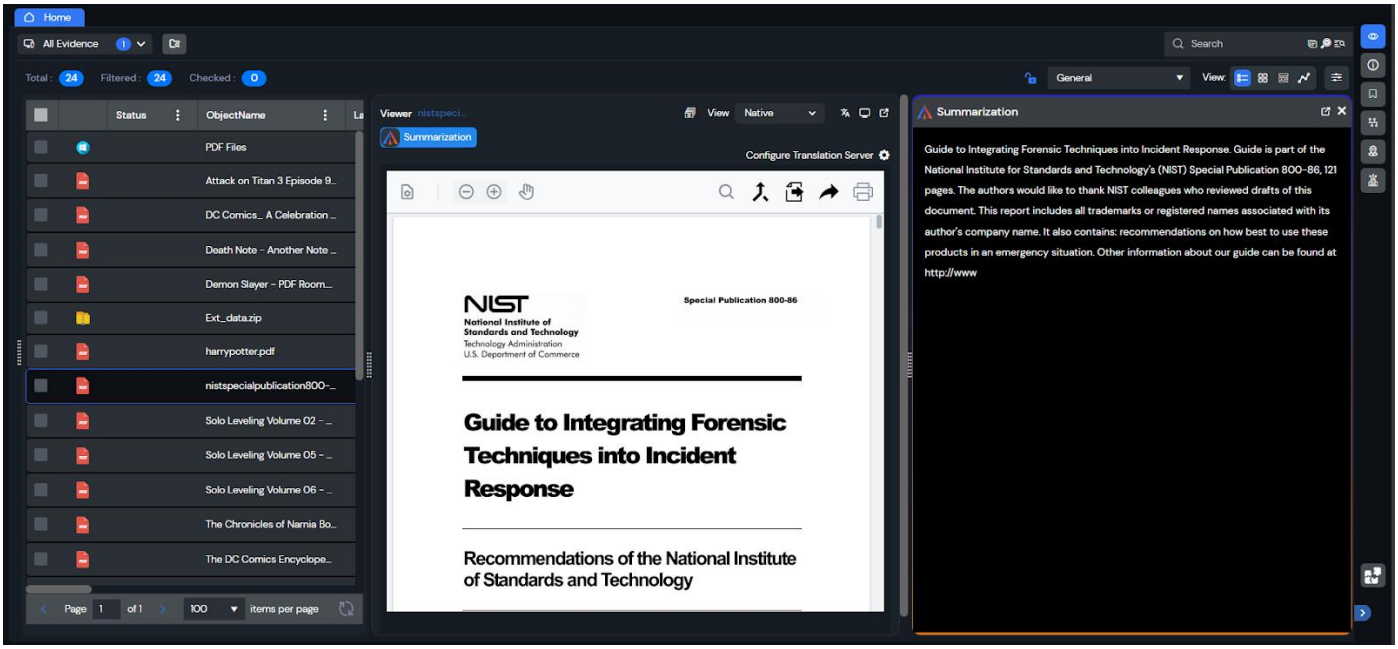
- **Time Taken**

Hardware	Time Taken
CPU	1 minute 4 sec
GPU	1 minute 46 sec

The GPU outperforms the CPU in analytical depth and output quality, highlighting its superior processing capability.

3.2.1 CPU Results (Summary)

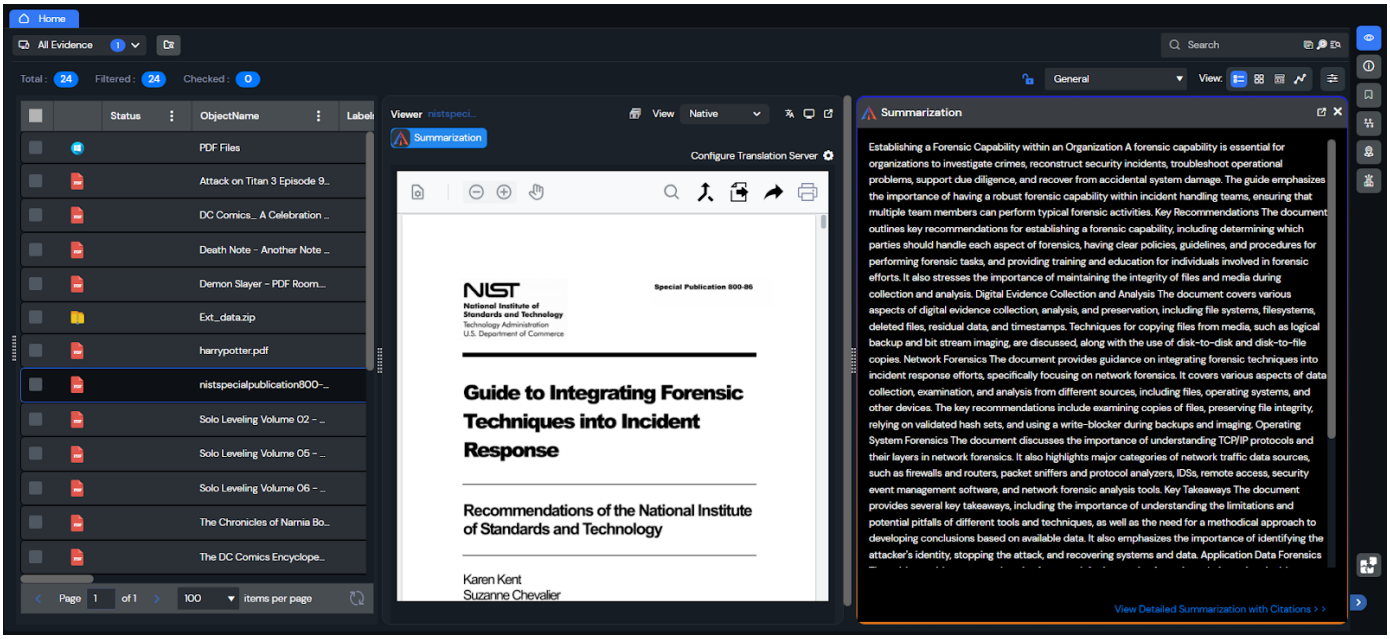
Guide to Integrating Forensic Techniques into Incident Response. Guide is part of the National Institute for Standards and Technology's (NIST) Special Publication 800-86, 121 pages. The authors would like to thank NIST colleagues who reviewed drafts of this document. This report includes all trademarks or registered names associated with its author's company name. It also contains: recommendations on how best to use these products in an emergency situation. Other information about our guide can be found at http://www"



3.2.2 GPU Results (Summary)

Establishing a Forensic Capability within an Organization. A forensic capability is essential for organizations to investigate crimes, reconstruct security incidents, troubleshoot operational problems, support due diligence, and recover from accidental system damage. The guide emphasizes the importance of having a robust forensic capability within incident handling teams, ensuring that multiple team members can perform typical forensic activities. Key Recommendations. The document outlines key recommendations for establishing a forensic capability, including determining which parties should handle each aspect of forensics, having clear policies, guidelines, and procedures for performing forensic tasks, and providing training and education for individuals involved in forensic efforts. It also stresses the importance of maintaining the integrity of files and media during collection and analysis. Digital Evidence Collection and Analysis. The document covers various aspects of digital evidence collection, analysis, and preservation, including file systems, filesystems, deleted files, residual data, and timestamps. Techniques for copying files from media, such as logical backup and bit stream imaging, are discussed, along with the use of disk-to-disk and disk-to-file copies. Network Forensics. The document provides guidance on integrating forensic techniques into incident response efforts, specifically focusing on network forensics. It covers various aspects of data collection, examination, and analysis from different sources, including files, operating systems, and other devices. The key recommendations include examining copies of files, preserving file integrity, relying on validated hash sets, and using a write-blocker during backups and imaging.. Operating System Forensics. The document discusses the importance of understanding TCP/IP protocols and their layers in network forensics. It also highlights major categories of network traffic data sources, such as firewalls and routers, packet sniffers and protocol analyzers, IDSs, remote access, security event management software, and network forensic analysis tools. Key Takeaways. The document provides several key takeaways, including the importance of understanding the limitations and potential pitfalls of different tools and techniques, as well as the need for a methodical approach to developing conclusions based on available data. It also emphasizes the importance of identifying the attacker's identity, stopping the attack, and recovering systems and data. Application Data Forensics. The guide provides a comprehensive framework for integrating forensic techniques into incident response efforts, specifically focusing on applications and application-related data. It covers various aspects of collecting, examining, and analyzing application data from different sources such as file systems, volatile OS data, network traffic, and logs. Incident Response and Forensic Techniques. The document provides an overview of integrating forensic techniques into incident response, focusing on data acquisition, analysis, and preservation. It covers various aspects such as data concealment, memory management, and network forensic analysis tools. The guide also discusses incident

handling, including containment, exercises, and outsourcing, as well as information system life cycles and digital evidence management. Overall, the document provides a comprehensive framework for integrating forensic techniques into incident response efforts, emphasizing the importance of understanding the limitations and potential pitfalls of different tools and techniques.



Contact Exterro

If you have any questions, please refer to this document, or any other related materials provided to you by Exterro. For usage questions, please check with your organization's internal application administrator. Alternatively, you may contact your Exterro Training Manager or other Exterro account contact directly.

For technical difficulties, support is available through support@exterro.com.

Contact:

Exterro, Inc.

2175 NW Raleigh St., Suite 110

Portland, OR 97210.

Telephone: 503-501-5100

Toll Free: 1-877-EXTERRO (1-877-398-3776)

Fax: 1-866-408-7310

General E-mail:info@exterro.com

Website: www.exterro.com