White Paper

# ARBOR Technology - AI Workload Guidance

## Collaboration with Gimlet Labs

www.arbor-technology.com

## Summary

ARBOR Technology provides a diverse range of edge hardware solutions tailored for industries such as retail, transportation, industrial automation, and medical. This analysis aims to deliver customer-facing guidance for selecting the right hardware based on specific AI application needs.

In collaboration with Gimlet Labs, an ARBOR Technology partner specializing in AI software, this analysis offers device recommendations informed by AI workload sizing and operational constraints. Industry-specific workload examples are included, along with estimates of the AI capacity for each device.

Please note that these figures are approximate and may vary depending on the specific application and any additional workloads running on the device. For a more detailed assessment of your AI application, we encourage you to consult with ARBOR Technology and Gimlet Labs.

## ARBOR Products Included in Analysis

### Retail (Signage Players and Display Terminals)

| Product | Description |
|---|---|
| IEC-3366 | Digital Signage Player with Intel® 11th Core / Celeron® Processor |
| IEC-3702 | Digital Signage Player with Intel® 11th Core i5-1135G7 Processor |
| ELIT-1060 | Ultra Compact Triple Display Terminal Powered by Intel® Elkhart Lake Processor |
| IEC-3904-1145G7E System R1.0 | Digital Signage Player with Intel® 11th Gen. Core™ i7/i5/i3/ Celeron processor |

### Transportation

| Product | Description |
|---|---|
| ARES-5320 | Fanless DIN-Rail Embedded System with Intel® Elkhart Lake Atom™ Processor |
| FPC-5211 | Intel® 14th / 13th Gen. Core™ Fanless Edge AI Computer supporting NVIDIA® RTX A2000 GPU |

### Industrial Automation

| Product | Description |
|---|---|
| ARES-1980 | Fanless Rugged Controller with 11th Gen. Intel® Core™ i Processor (Tiger Lake UP3) |
| ARES-1983H | Machine Vision Controller with Intel® 14th / 13th / 12th Gen. Core™ i7/i5/i3 Processor |
| FPC-821X | Robust Box PC with Intel® 14th/ 13th/ 12th Generation Core™ i9/i7/i5/i3 Processor |
| iTC-1150R R1.0 | 15" Industrial Panel PC |
| SB-244-1J64 | Industrial Fanless PC w/ Intel® Celeron® J6412 CPU Elkhart Lake Processor |

### Medical Imaging

| Product | Description |
|---|---|

| | |
|---|---|
| AEC-6100 | NVIDIA® Jetson AGX Orin™ AI Embedded Computer Supporting GMSL Camera |
| FPC-9108-P6-G3 | Ruggedized Edge AI Computing Platform Supporting NVIDIA® GeForce RTX 3090 GPU, Intel ®10th Gen Xeon® Core™ Processor with 6 GbE PoE |

## AI Workloads Included in Analysis

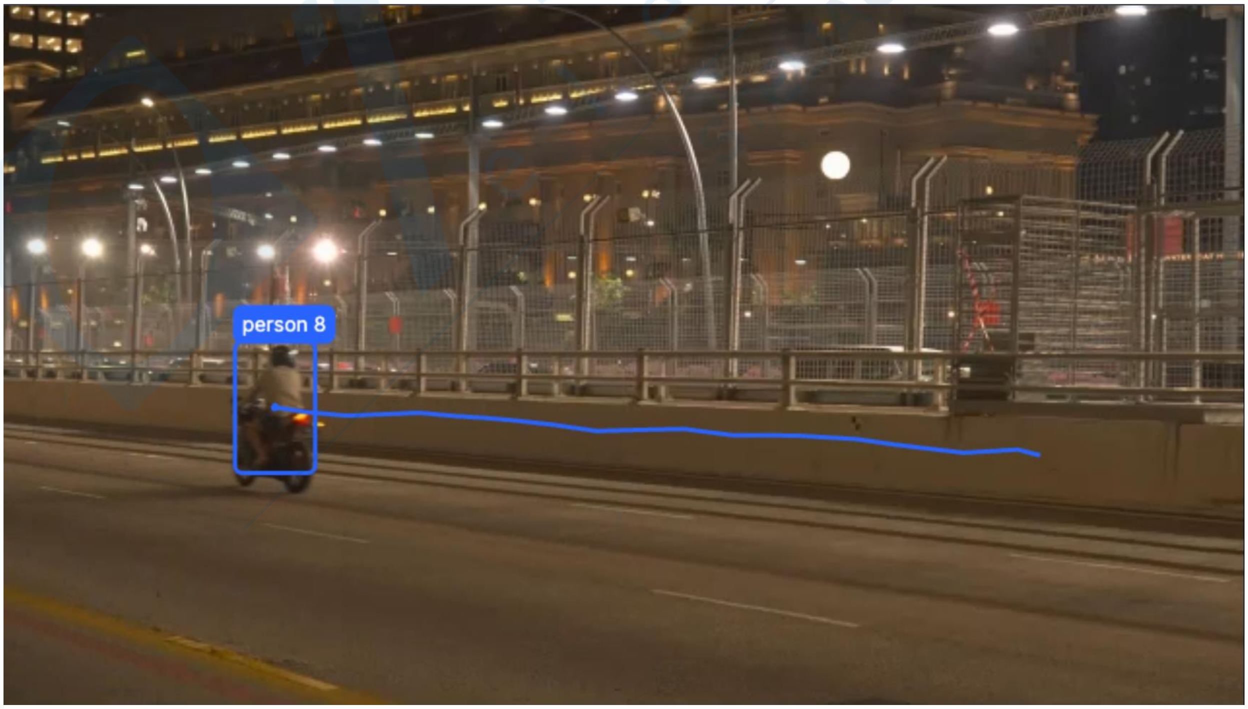
There were four primary workload categories included in this analysis:

1. Vision: Object detection, tracking, counting
2. Vision: Semantic segmentation
3. Multimodal: Vision language models
4. Text: Language models

Each of these has different applications depending on the target industry. Example applications will be covered in the industry analysis sections ahead.

### Vision: Object detection, tracking, counting

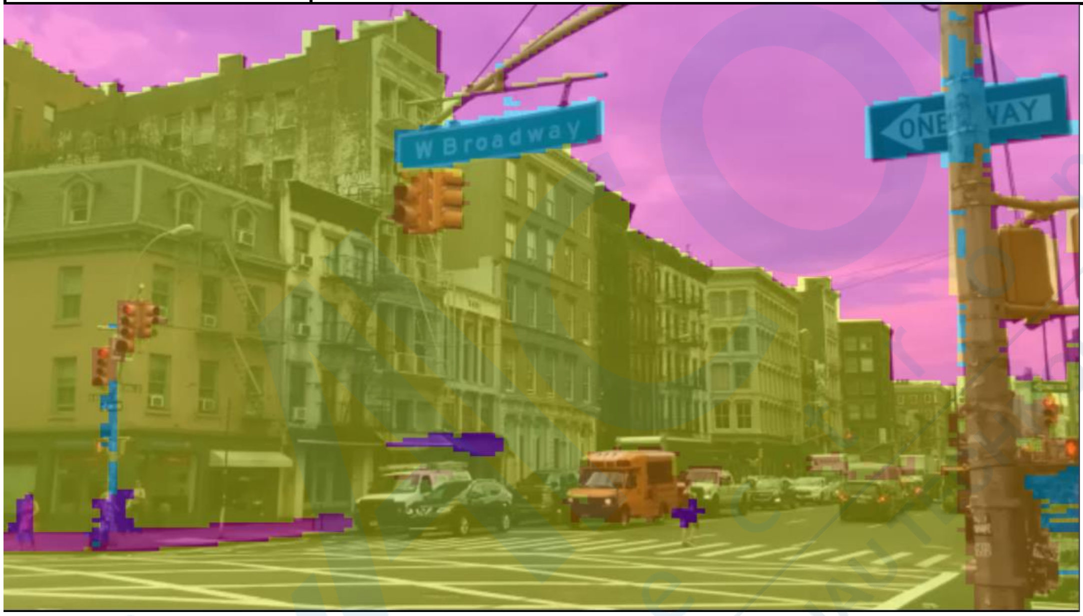
| | |
|---|---|
| **Purpose** | Detect, track, and count objects in video or image data |
| **Common models** | YOLO, Faster-RCNN |
| **Target Metric** | Frames per second ("Real time" is considered ~10-15 FPS) |
| **Workload size** | Small to moderate |



*Example Object Detection, Tracking, and Counting workload: Human path tracking (Screenshot from Gimlet AI platform)*

## Vision: Semantic segmentation

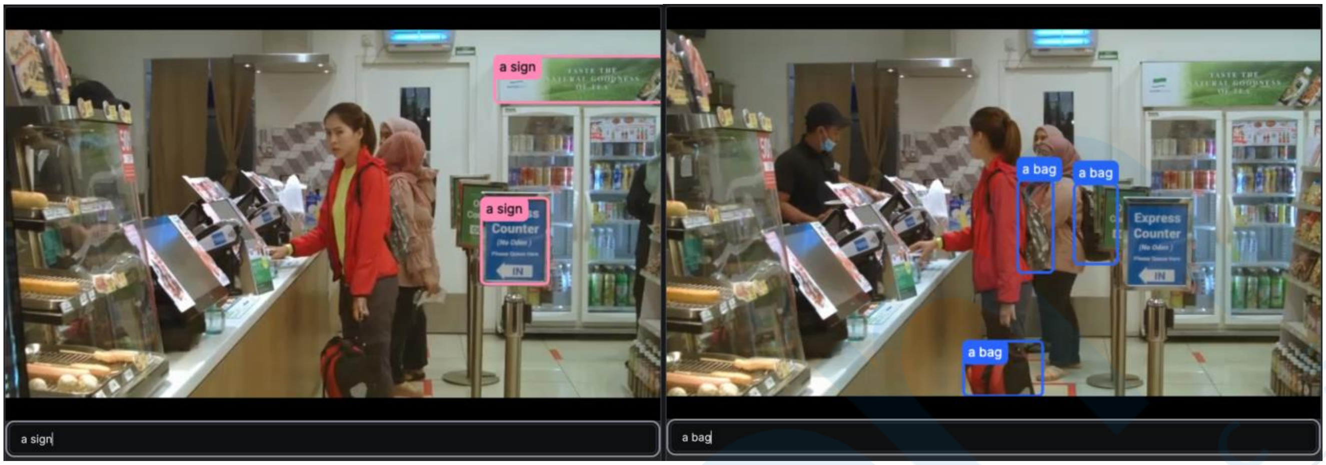
| | |
|---|---|
| **Purpose** | Fine-grained segmentation of images based on semantic categories |
| **Common models** | Segformer, U-Net |
| **Target Metric** | Frames per second (“Real time” performance is considered ~10-15 FPS) |
| **Workload size** | Moderate |



*Example Semantic Segmentation workload: Outdoor landscape segmentation (Screenshot from Gimlet AI platform)*

## Multimodal: Vision Language Models

| | |
|---|---|
| **Purpose** | Detection or segmentation of video/image data based on text prompts, which can be provided on the fly (rather than precoded) |
| **Common models** | OWL-ViT, LLaVA |
| **Target Metric** | Frames per second (“Real time” performance is considered ~10-15 FPS) |
| **Workload size** | Moderate to large |



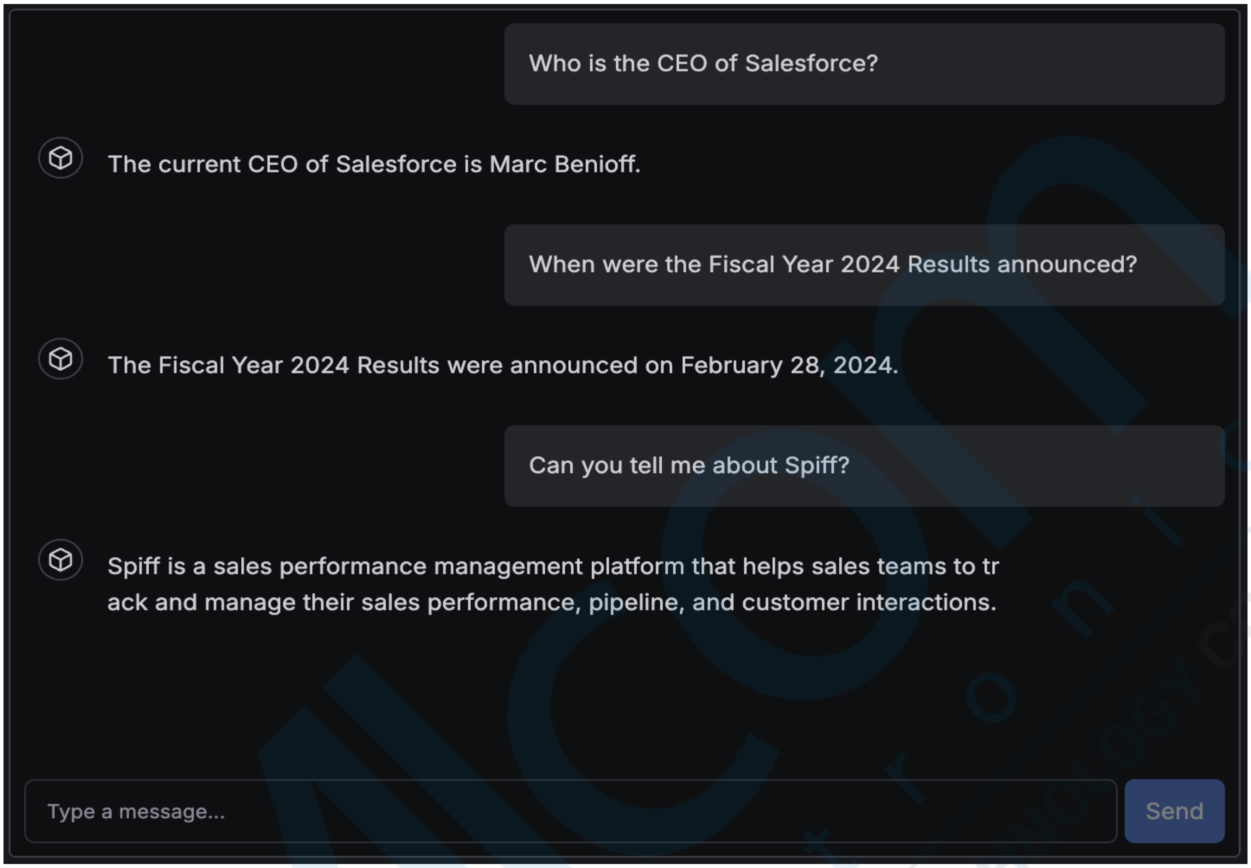
a sign

a bag

*Example Vision Language Model workload: Dynamic retail analytics*
*Note the prompts at the bottom ("a sign", "a bag") can be specified in real time without retraining.*
*(Screenshot from Gimlet AI platform)*

## Text: Language Models

| | |
|---|---|
| **Purpose** | Text processing models for tasks such as question answering on specific corpuses of data, or other text-based interaction |
| **Common models** | Llama, Mixtral |
| **Target Metric** | Tokens per second (Words are mapped to "tokens". For example, imagine a 100 word input prompt to a language model, which generates a 100 word response. At a rate of 100 Tokens per second, we would expect the result to take about 2 seconds to compute. |
| **Workload size** | Large to very large |



Who is the CEO of Salesforce?

The current CEO of Salesforce is Marc Benioff.

When were the Fiscal Year 2024 Results announced?

The Fiscal Year 2024 Results were announced on February 28, 2024.

Can you tell me about Spiff?

Spiff is a sales performance management platform that helps sales teams to tr
ack and manage their sales performance, pipeline, and customer interactions.

*Example Language Model workload: Question answering and summarization on a private dataset*
*Example Dataset: Salesforce press releases*
*(Screenshot from Gimlet AI platform)*

## Methodology

This analysis estimates performance based on AI benchmarks on similar devices. It is intended to be a high-level projection to understand approximate performance. Estimates include interpolation based on device capabilities. Real world performance will vary based on the exact configuration of the device and the specific workload. More detailed and specific estimates are available upon request if the workload and exact device configuration are specified.

## Analysis: Retail Market

For the retail market, the analysis focused on ARBOR's line of digital signage players and display terminals. These devices are purely Intel-based, with the exception of the ELIT-1060, which has a Hailo accelerator.

| Product | Description |
|---|---|
| IEC-3366 | Digital Signage Player with Intel® 11th Core / Celeron® Processor |
| IEC-3702 | Digital Signage Player with Intel® 11th Core i5-1135G7 Processor |
| ELIT-1060 | Ultra Compact Triple Display Terminal Powered by Intel® Elkhart Lake Processor |
| IEC-3904-1145G7E System R1.0 | Digital Signage Player with Intel® 11th Gen. Core™ i7/i5/i3/ Celeron processor |

*The IEC-3366 is pictured on the left and the ELIT-1060 is pictured on the right.*

The specs of these devices make them suited to object detection and semantic segmentation workloads. Vision language models and language models would be best suited for devices with NVIDIA graphics, such as the AEC-6100 or the FPC line.

## Use cases by AI workload type

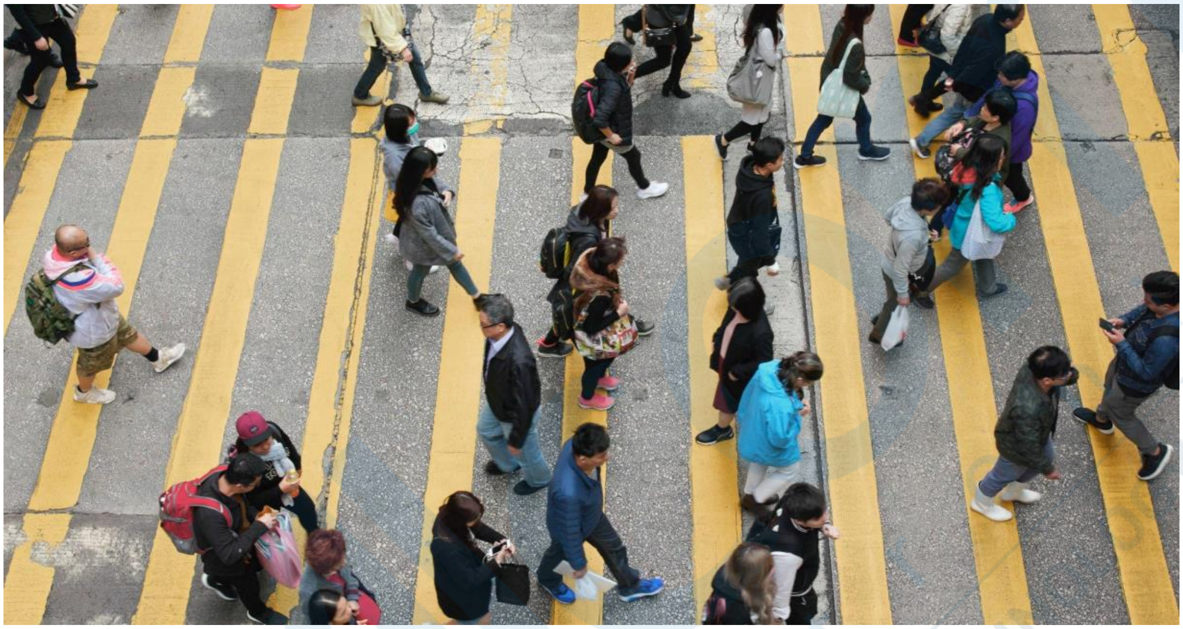
In the retail space, focused on digital signage/display terminals specifically, here are some examples use cases in the object detection and semantic segmentation workload categories.

### Object detection, tracking, counting

- Demographic profiling of people (age, gender, etc)
- Person tracking, detection, counting
- Gaze detection and tracking
- Identify clothing items and style
- Crowd and line identification

Semantic segmentation

- Scenery segmentation for context awareness
- Crowd flow and human traffic analysis
- Identify weather and ambient conditions



*Use case example: Crowd flow analytics for retail environments (Image Source: Adobe Stock)*

*Use case example: Gaze, gender, and clothing detection for smart billboards*

## Estimated performance by device

The performance for each workload category on the device is based on benchmarks collected from the Gimlet platform. These estimates are designed to provide a realistic view of typical performance for a "typical" workload. However, actual performance can vary due to several factors, including:

- **Device Specifications**: For instance, an Intel i3 and i7 within the same generation will deliver different levels of performance.
- **Camera Resolution and Type**: Higher resolution or specialized camera types may impact processing demands.
- **Preprocessing and Postprocessing Steps**: Variations in data preparation or output handling can influence performance.
- **Model Size**: Models often come in multiple sizes; for example, YOLO ranges from “nano” to “x-large.”

To help account for these variations, we provide a range of estimates to include both “lower-end” and “higher-end” workloads within each category. For example, a higher-end workload for object detection might involve a larger model combined with a high-resolution camera feed. This approach offers flexibility and guidance for understanding potential performance across a range of scenarios.

### Object detection, tracking, and counting

| **Device** | **Estimated AI latency per frame** | **Estimated Frames per second** |
|---|---|---|
| *IEC-3366 (i5)* | 50 - 100 ms | 10 - 20 FPS |
| *IEC-3702* | 50 - 100 ms | 10 - 20 FPS |
| *IEC-3904-1145G7E System R1.0 (i5)* | 50 - 100 ms | 10 - 20 FPS |
| *ELIT-1060 (Hailo)* | 20 - 40 ms | 25 - 50 FPS |

### Semantic segmentation

| **Device** | **Estimated AI latency per frame** | **Estimated Frames per second** |
|---|---|---|
| *IEC-3366 (i5)* | 300 ms - 1 s | 1 - 3.3 FPS |
| *IEC-3702* | 300 ms - 1 s | 1 - 3.3 FPS |
| *IEC-3904-1145G7E System R1.0 (i5)* | 300 ms - 1 s | 1 - 3.3 FPS |
| *ELIT-1060 (Hailo)* | 50 ms | 20 FPS |

## Discussion

All four devices have good AI capabilities for retail use cases in object detection, tracking, and counting. Because the different devices have different configurations, we assumed a Core i5 configuration for the IEC-3366 and IEC-3904-1145G7E devices. Note that estimated performance will vary based on the exact configuration, so if higher performance is desired, a Core i7 or i9 would be recommended for those devices.

The non-Hailo devices (IEC-366, IEC-3702, and IEC-3904-1145G7E) can support between 1-2 real time camera streams for the object detection, tracking, and counting workloads, depending on the workload complexity. The Hailo device (ELIT-1060) can support between 2-5 real time camera streams, depending on the workload complexity, when configured with the Hailo accelerator.

As for semantic segmentation, the ELIT-1060 with Hailo is recommended if real time performance (10 FPS) is required. However, for use cases where 1-3 FPS is permissible, then the Intel-only configurations are viable options for those use cases.

## Analysis: Transportation Market

For the transportation market, the analysis focused on two of ARBOR's embedded devices. Both devices support Intel-based execution, and one device supports the option to add NVIDIA RTX A2000.

| Product | Description |
|---|---|
| ARES-5320 | Fanless DIN-Rail Embedded System with Intel® Elkhart Lake Atom™ Processor |
| FPC-5211 | Intel® 14th / 13th Gen. Core™ Fanless Edge AI Computer supporting NVIDIA® RTX A2000 GPU |

*The ARES-5320 is pictured on the left and the FPC-5211 is pictured on the right.*

The FPC-5211 with NVIDIA graphics, due to its high AI compute capacity, would be well suited for most AI applications, including vision language models and language models. The ARES-5320 would be better suited for simple object detection workloads.

## Use cases by AI workload type

In the transportation space, here are some prospective use cases in each category.

### Object detection, tracking, counting

- Driver/conductor alertness detection
- Vehicle capacity utilization detection
- Weight estimation of car/train/cabin
- Human fall detection
- Animal intrusion detection
- Fare evasion detection

### Semantic segmentation

- Railway or roadway inspection

### Vision language models

- Dynamic analytics and video search (“identify people smoking on the train”)
- Lost item detection
- Analyze changes in passenger or traffic behavior
- Video summarization

### Language models

- Question answering for operators about service or capacity
- Question answering for passengers about train stops, bus stops, and other questions

*Use case example: Fare evasion identification*
*(Image Source: Adobe Stock)*

*Use case example: Alerting for animal intrusion on railway racks*

## Estimated performance by device

The performance for each workload category on the device is based on benchmarks collected from the Gimlet platform. These estimates are designed to provide a realistic view of typical performance for a "typical" workload. However, actual performance can vary due to several factors, including:

- **Device Specifications**: For instance, an Intel i3 and i7 within the same generation will deliver different levels of performance.
- **Camera Resolution and Type**: Higher resolution or specialized camera types may impact processing demands.
- **Preprocessing and Postprocessing Steps**: Variations in data preparation or output handling can influence performance.
- **Model Size**: Models often come in multiple sizes; for example, YOLO ranges from “nano” to “x-large.”

To help account for these variations, we provide a range of estimates to include both “lower-end” and “higher-end” workloads within each category. For example, a higher-end workload for object detection might involve a larger model combined with a high-resolution camera feed. This approach offers flexibility and guidance for understanding potential performance across a range of scenarios.

### Object detection, tracking, and counting

| **Device** | **Estimated AI latency per frame** | **Estimated Frames per second** |
|---|---|---|
| *ARES-5320* | 400 ms - 2 s | 0.5 - 2.5 FPS |
| *FPC-5211 (i5, no RTX)* | 40 - 80 ms | 12.5 - 25 FPS |
| *FPC-5211 (with RTX)* | 5 - 10 ms | 100 - 200 FPS |

### Semantic segmentation

| **Device** | **Estimated AI latency per frame** | **Estimated Frames per second** |
|---|---|---|
| *FPC-5211 (i5, no RTX)* | 250 - 500 ms | 2 - 4 FPS |
| *FPC-5211 (with RTX)* | 10 - 20 ms | 50 - 100 FPS |

### Vision language models

| Device | Estimated AI latency per frame | Estimated Frames per second |
|---|---|---|
| *FPC-5211 (with RTX)* | 10 - 20 ms | 50 - 100 FPS |

### Language models

For the language models, we assume input prompts of ~100 tokens (which is about 100 words) and output responses of about that same size. The size of inputs/outputs for language models will depend on the prompt and the topic.

| Device | Estimated total response latency | Estimated latency to first token |
|---|---|---|
| *FPC-5211 (with RTX)* | 500 ms - 1 s | 250 - 500 ms |

## Discussion

Among these devices, we can see a broad spectrum of AI capabilities. The ARES-5320 provides options for lower end object detection, counting, and tracking workloads where real time performance is not crucial. For the transportation industry, the ARES-5320 could be a good fit for analytics use cases such as periodic sampling of crowds and traffic flow, train capacity monitoring, and vehicle arrival detection.

The FPC-5211 provides two tiers of AI capabilities depending on the needs of the use case. The base configuration without NVIDIA graphics can support real-time performance for object detection, tracking, and counting use cases on 1-2 camera streams. It is also capable of non-real-time but still performant semantic segmentation use cases.

With the NVIDIA RTX A2000 added on, the capabilities extend even further. Use cases involving vision language models and language models are unlocked. Real time performance is supported for 10-20 cameras for object detection, counting, and tracking. Real time performance is supported for 2-10 cameras for semantic segmentation workloads and 5-10 cameras for vision language model workloads. The FPC-5211 with the NVIDIA RTX A2000 also unlocks language model workloads. Interactive language model performance is available, with sub 1s response latencies expected for simple queries.

## Analysis: Industrial automation

For the industrial automation market, the analysis focused on five of ARBOR’s devices. These devices have a mix of capabilities, offering CPUs between the Intel Celeron and Intel Core i7 models, some with integrated graphics, and the ARES-1980 offering the option for a Hailo AI accelerator.

| **Product** | **Description** |
|---|---|
| ARES-1980 | Fanless Rugged Controller with 11th Gen. Intel® Core™ i Processor (Tiger Lake UP3) |
| ARES-1983H | Machine Vision Controller with Intel® 14th / 13th / 12th Gen. Core™ i7/i5/i3 Processor |
| FPC-821X | Robust Box PC with Intel® 14th/ 13th/ 12th Generation Core™ i9/i7/i5/i3 Processor |
| iTC-1150R R1.0 | 15" Industrial Panel PC |
| SB-244-1J64 | Industrial Fanless PC w/ Intel® Celeron® J6412 CPU Elkhart Lake Processor |

*The FPC-821X is pictured on the left and the SB-244-1J64 is pictured on the right.*

This group of devices would be best suited to embedded vision workloads, including object detection and semantic segmentation.

### Use cases by AI workload type

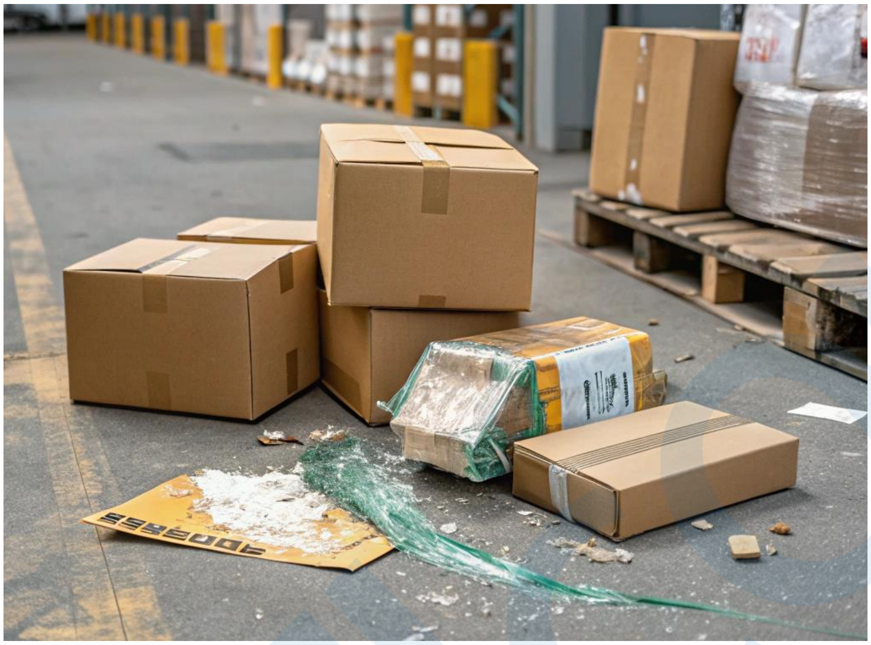
In the industrial automation space, here are some prospective use cases in each category.

#### Object detection, tracking, counting

- Inventory counting/tracking
- Worker task identification and analytics
- Detection of specific SKUs

Semantic segmentation

- Defect detection



*Use case example: Identifying damage to products and/or boxes*

*Use case example: Worker task identification and analytics*

## Estimated performance by device

The performance for each workload category on the device is based on benchmarks collected from the Gimlet platform. These estimates are designed to provide a realistic view of typical performance for a "typical" workload. However, actual performance can vary due to several factors, including:

- **Device Specifications**: For instance, an Intel i3 and i7 within the same generation will deliver different levels of performance.
- **Camera Resolution and Type**: Higher resolution or specialized camera types may impact processing demands.
- **Preprocessing and Postprocessing Steps**: Variations in data preparation or output handling can influence performance.
- **Model Size**: Models often come in multiple sizes; for example, YOLO ranges from “nano” to “x-large.”

To help account for these variations, we provide a range of estimates to include both “lower-end” and “higher-end” workloads within each category. For example, a higher-end workload for object detection might involve a larger model combined with a high-resolution camera feed. This approach offers flexibility and guidance for understanding potential performance across a range of scenarios.

### Object detection, tracking, and counting

| **Device** | **Estimated AI latency per frame** | **Estimated Frames per second** |
|---|---|---|
| *ARES-1980 (i5, no Hailo)* | 50 - 100 ms | 10 - 20 FPS |
| *ARES-1980 (with Hailo)* | 20 - 40 ms | 25 - 50 FPS |
| *ARES-1983 (i5)* | 40 - 80 ms | 12.5 - 25 FPS |
| *FPC-821X* | 40 - 80 ms | 12.5 - 25 FPS |
| *iTC-1150R R1.0* | 100 - 200 ms | 5 - 10 FPS |
| *SB-244-1J64* | 100 - 200 ms | 5 - 10 FPS |

### Semantic segmentation

| **Device** | **Estimated AI latency per frame** | **Estimated Frames per second** |
|---|---|---|
| *ARES-1980 (no Hailo)* | 300 - 600 ms | 1.6 - 3.3 FPS |
| *ARES-1980 (with Hailo)* | 50 - 100 ms | 10 - 20 FPS |

| | | |
|---|---|---|
| *ARES-1983* | 250 - 500 ms | 2 - 4 FPS |
| *FPC-821X* | 250 - 500 ms | 2 - 4 FPS |
| *iTC-1150R R1.0* | 400 - 800 ms | 1.2 - 2.5 FPS |
| *SB-244-1J64* | 400 - 800 ms | 1.2 - 2.5 FPS |

## Discussion

The selected ARBOR devices offer a variety of packaging options for AI vision workloads in the industrial automation space. The top performance in both object detection, tracking, and counting workloads as well as semantic segmentation comes from the ARES-1980 configured with the Hailo, which can support real time camera streams for 1-5 cameras depending on the workload.

The other devices (ARES-1983, FPC-821X, iTC-1150R R1.0, SB-244-1J64, as well as ARES-1980 without Hailo) also offer real time or close to real time object detection, tracking, and counting performance for 1-2 cameras on pure Intel hardware. As for semantic segmentation workloads, there is still viable performance on Intel-only hardware as long as the use case does not require fully real time segmentation capabilities (which would be 10-15 FPS). For inspection and defect detection use cases where the camera receives an input of a still image on the order of once every second, these could be great options.

## Analysis: Medical Imaging

For the medical imaging market, the analysis focused on two of ARBOR's embedded devices. Both devices support high-level NVIDIA graphics, with the FPC-9108-P6-G3 also offering Xeon Core processors in addition.

| **Product** | **Description** |
|---|---|
| AEC-6100 | NVIDIA® Jetson AGX Orin™ AI Embedded Computer Supporting GMSL Camera |
| FPC-9108-P6-G3 | Ruggedized Edge AI Computing Platform Supporting NVIDIA® GeForce RTX 3090 GPU, Intel ®10th Gen Xeon® Core™ Processor with 6 GbE PoE |

*The AEC-6100 is pictured on the left and the FPC-9108-P6-G3 is pictured on the right.*

Both devices can support different types of AI workloads, with the FPC-9108-P6-G3 supporting a higher level of performance and workload sophistication.

## Use cases by AI workload type

In the medical imaging space, here are some prospective use cases in each category.

### Object detection, tracking, counting

- Volumetric CNNs (for processing 3D imaging data)
- Tracking, counting, and identifying biological sample components such as cells

### Semantic segmentation

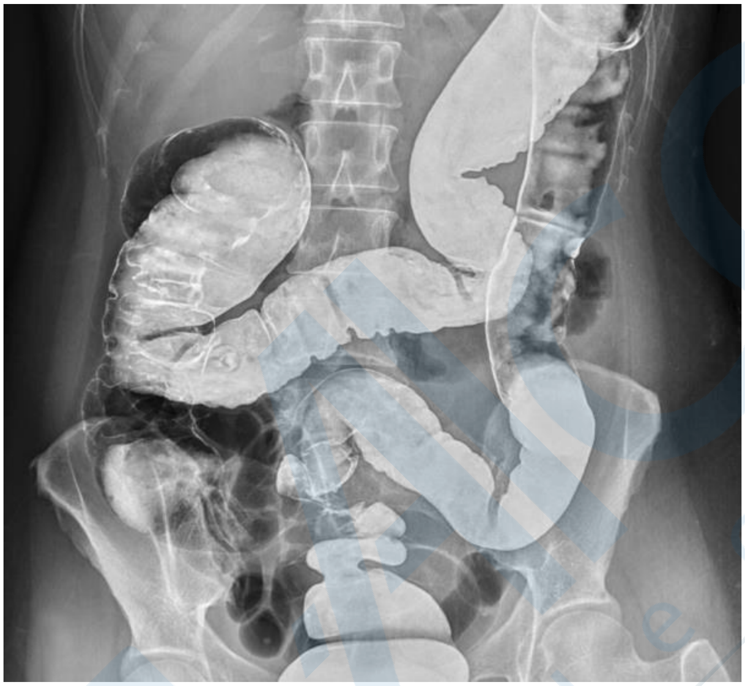
- Tumor segmentation
- Organ delineation
- Size measurement and area calculation of different biological components in samples and imaging data

### Vision language models

- Biological sample text description / label generation
- Description / label generation of medical imaging
- Open-ended visual search across medical imaging data and records

### Language models

- Question answering about biological samples
- Provider note taking and transcript summarization



*Use case example: Segmentation of organs and tissues*

*Use case example: Counting and identifying benign vs malignant cells*

## Estimated performance by device

The performance for each workload category on the device is based on benchmarks collected from the Gimlet platform. These estimates are designed to provide a realistic view of typical performance for a "typical" workload. However, actual performance can vary due to several factors, including:

- **Device Specifications**: For instance, an Intel i3 and i7 within the same generation will deliver different levels of performance.
- **Camera Resolution and Type**: Higher resolution or specialized camera types may impact processing demands.
- **Preprocessing and Postprocessing Steps**: Variations in data preparation or output handling can influence performance.
- **Model Size**: Models often come in multiple sizes; for example, YOLO ranges from “nano” to “x-large.”

To help account for these variations, we provide a range of estimates to include both “lower-end” and “higher-end” workloads within each category. For example, a higher-end workload for object detection might involve a larger model combined with a high-resolution camera feed. This approach offers flexibility and guidance for understanding potential performance across a range of scenarios.

### Object detection, tracking, and counting

| Device | Estimated AI latency per frame | Estimated Frames per second |
|---|---|---|
| *AEC-6100* | 20 - 40 ms | 25 - 50 FPS |
| *FPC-9108-P6-G3* | 5 - 10 ms | 100 - 200 FPS |

### Semantic segmentation

| Device | Estimated AI latency per frame | Estimated Frames per second |
|---|---|---|
| *AEC-6100* | 50 - 100 ms | 10 - 20 FPS |
| *FPC-9108-P6-G3* | 10 - 20 ms | 50 - 100 FPS |

### Vision language models

| Device | Estimated AI latency per frame | Estimated Frames per second |
|---|---|---|
| *AEC-6100* | 40 - 80 ms | 12.5 - 25 FPS |
| *FPC-9108-P6-G3* | 10 - 20 ms | 50 - 100 FPS |

### Language models

For the language models, we assume input prompts of ~100 tokens (which is about 100 words) and output responses of about that same size. The size of inputs/outputs for language models will depend on the prompt and the topic.

| Device | Estimated total response latency | Estimated latency to first token |
|---|---|---|
| *AEC-6100* | 1.2 - 5 s | 600 ms - 2.5 s |
| *FPC-9108-P6-G3* | 200 - 400 ms | 100 - 200 ms |

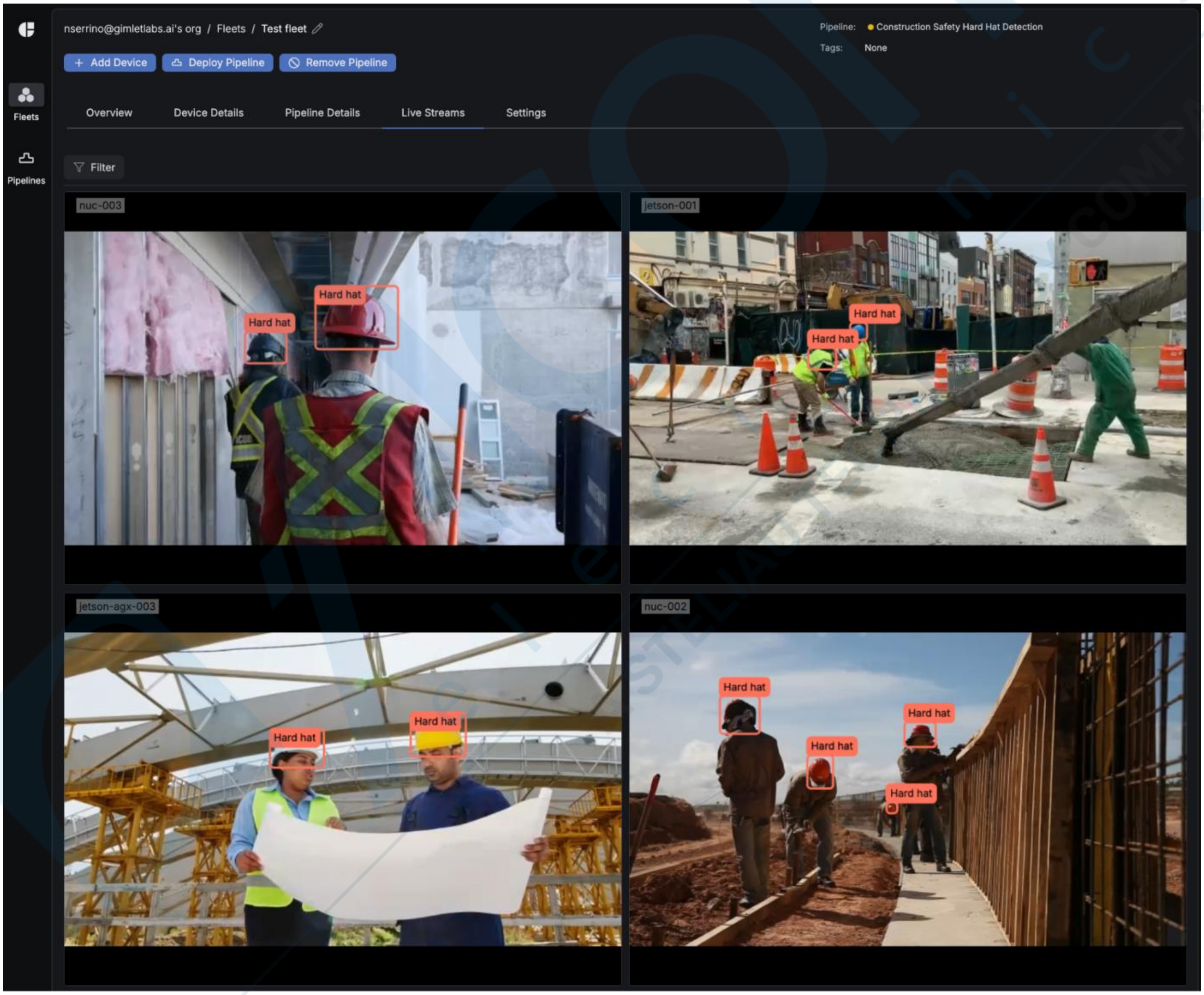
## Discussion

Both the AEC-6100 and the FPC-9108-P6-G3 support a spectrum of AI workloads, including vision language models and language models. The FPC-9108-P6-G3 offers a higher performance across all workload types, supporting between 5-20 real time camera streams for vision workloads and high responsiveness for language workloads. The AEC-6100 offers compelling performance for a smaller number of cameras across vision workloads, from 1-5 cameras depending on the workload complexity. It also has the capability to run language models at a slower but still solid performance than the FPC-9108-P6-G3.

## About Gimlet Labs

Gimlet Labs is an AI software company specializing in the deployment of AI workloads across edge and cloud systems. The Gimlet developer platform simplifies AI model deployment across diverse devices, eliminating the need for manual porting. It supports a wide range of hardware, including NVIDIA, Intel, AMD and more.

Once deployed, the platform offers robust monitoring and fleet management capabilities, enabling seamless oversight of AI operations.



nserrino@gimletlabs.ai's org / Fleets / Test fleet

Pipeline: Construction Safety Hard Hat Detection

Tags: None

Screenshot of Gimlet AI platform running *Construction safety monitoring edge AI workload*