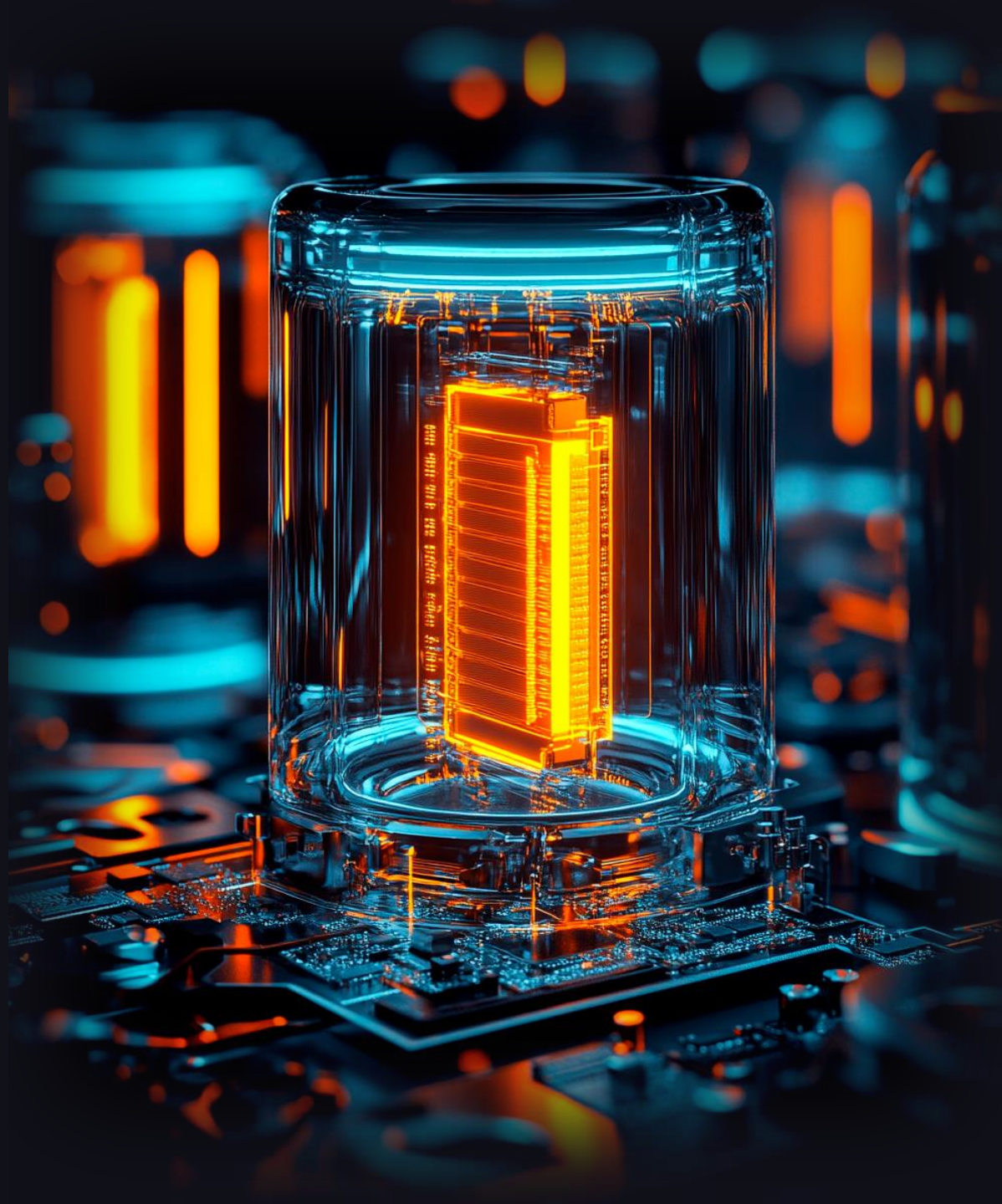


# РЕШЕНИЯ ДЛЯ УСКОРЕНИЯ РАБОТЫ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

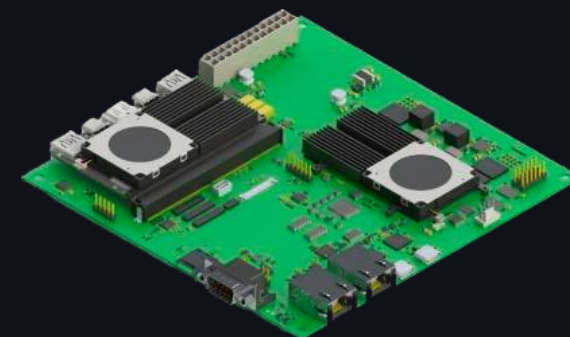
Каталог ИИ-решений



# ОТЕЧЕСТВЕННЫЕ ВЫЧИСЛИТЕЛИ ДЛЯ ЗАПУСКА НЕЙРОННЫХ СЕТЕЙ

Вычислители предназначены для запуска ИИ-приложений и применяются в области периферийных вычислений, туманных вычислений, а также в стандартной инфраструктуре центров обработки данных.

На базе отечественного процессора собственной архитектуры и полного набора необходимого программного обеспечения.



# ТЕХНИЧЕСКИЕ ХАРАКТЕРИСТИКИ ПРОЦЕССОРА Н

**28 NM**

Техпроцесс

**16 МБ**

Встроенная память

**BGA 1296**

Корпус

**500-812 МГц**

Частота блока TPU

**≤ 25-30 Вт**

Энергопотребление

**MIPS64**

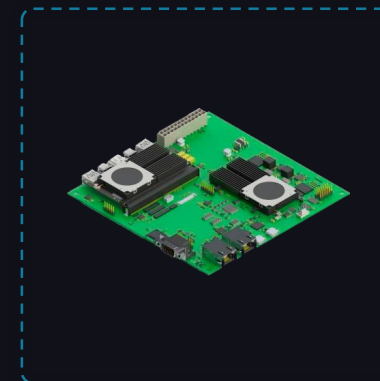
Управляющий  
процессор

**GEN3, ENDPOINT,  
8 ЛИНИЙ**

Контроллер PCIe

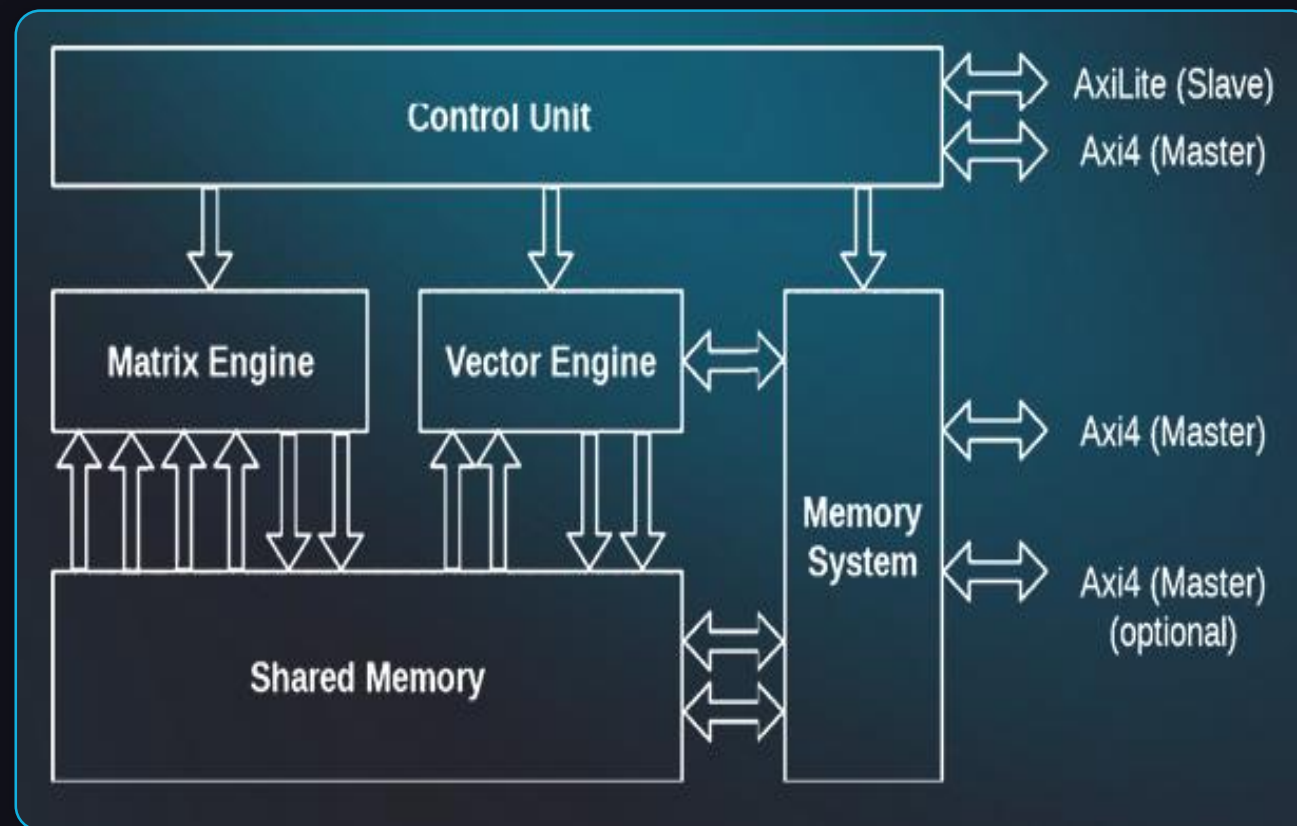
**2 КАНАЛА 72BIT  
ДО 32GB**

Контроллер DDR4+ECC



# СОБСТВЕННОЕ IP-ЯДРО

- ✓ Полностью российское AXI\_4 совместимое IP-ядро
- ✓ Расчет (inference) сверточных (CNN) и рекуррентных (RNN) нейронных сетей
- ✓ Вычислительных элементов 16K
- ✓ Максимальная рабочая частота 1 ГГц
- ✓ Тип данных int8
- ✓ Встроенная память до 16МБ
- ✓ Производительность 24 TOPs (int8) на систолическом массиве 128x128



# ЛИНЕЙКА РЕШЕНИЙ ДЛЯ ИНФЕРЕНСА

Характеристики	HP2	HPD2	HPQ2	НХ с сопроцессором общего назначения
Внешний вид				
Производительность пиковая	24 (int8) ТОПС	48 (int8) ТОПС	96 (int8) ТОПС	24 (int8) ТОПС
ResNet-50 (batch 1)	460 FPS	920 FPS	1840 FPS	460 FPS
ResNet-50 (batch 8)	1046 FPS	2092 FPS	4184 FPS	1046 FPS
Частота ускорителя тензорных вычислений	до 812 МГц	до 812 МГц	до 812 МГц	до 812 МГц
Интерфейсы	PCIe 2.0 x8 в Endpoint режиме	PCIe 3.0 x16 в Endpoint режиме	PCIe 3.0 x16 в Endpoint режиме	802.11a/b/g/n/ac; MIPI- SCI (2 канала); MIPI_DSI + TP; eDP; SDMMC; Audio codec
Форм-фактор	PCIe x16, 1 слот	PCIe x16, 1 слот	PCIe x16, 1 слот	Mini-ITX
Энергопотребление	20-23 Вт	<55 Вт	<105 Вт	<35 Вт

# ПЛАНЫ ПО РАЗВИТИЮ ПРОДУКТА



- ✓ densenet121
- ✓ efficientnet\_b0
- ✓ efficientnet\_b1
- ✓ efficientnet\_b2
- ✓ efficientnet\_b3
- ✓ inception\_v1
- ✓ inception\_v2
- ✓ inception\_v3
- ✓ inception\_v4
- ✓ Lenet5
- ✓ mobilenet\_v1
- ✓ mobilenet\_v2
- ✓ nasnet\_mobile
- ✓ pnasnet\_large
- ✓ pnasnet\_mobile
- ✓ resnet152
- ✓ resnet34
- ✓ resnet50
- ✓ resnet50\_mlperf
- ✓ resnet50\_v2
- ✓ Squeezenet
- ✓ ssd\_mobilenet\_v1
- ✓ ssd\_mobilenet\_v2
- ✓ scrfd 2.5g
- ✓ tiny\_yolo2
- ✓ tiny\_yolo3
- ✓ torch\_densenet169
- ✓ torch\_mobilenet\_v2
- ✓ torch\_resnet50
- ✓ vgg19
- ✓ xception
- ✓ yolo2
- ✓ yolo3
- ✓ yolo4
- ✓ Yolo5s
- ✓ ... и другие

# ПРОИЗВОДИТЕЛЬНОСТЬ (ИЗМЕРЕНИЯ НА ЧАСТОТЕ 750 МГц)

Сети (выборочно)		Batch	Модуль HP2		Модуль HPQ	
			Performance, FPS	Latency, ms	Performance, FPS	Latency, ms
1	Resnet-50 v1.5 MLperf	8	818,126	-	3272,504	-
		1	424,902	3.923	1699,608	3,923
2	Resnet-50	8	910,461	-	3641,844	-
		1	431,559	3.877	1726,236	3,877
3	Yolo2	8	-	-	-	-
		1	88,095	17.322	352,38	17,322
4	Yolo3	8	-	-	-	-
		1	31,315	42.712	125,26	42,712
5	Yolo4	8	-	-	-	-
		1	1,824	560.907	7,296	560,907
6	Yolo5s	8	-	-	-	-
		1	10.768	108,132	43,072	108,132
7	SSD Mobilenet v2	8	457,762	-	1831,048	-
		1	141,980	9,366	567,92	9,366
8	SCRFD 2,5g	8	110,823	-	443,292	-
		1	75,145	23,033	300,58	23,033

01

Разработка модели через стандартные фреймворки машинного обучения

02

Квантование модели

03

Формирование вычислительного графа

Строится вычислительное дерево и оптимизируется для выполнения на TPU

04

Расчет графа на TPU



# WEB-ИНТЕРФЕЙС ДЛЯ УДАЛЕННОЙ КОМПИЛЯЦИИ И КВАНТОВАНИЯ

## ПОДГОТОВКА К ИНФЕРЕНСУ:

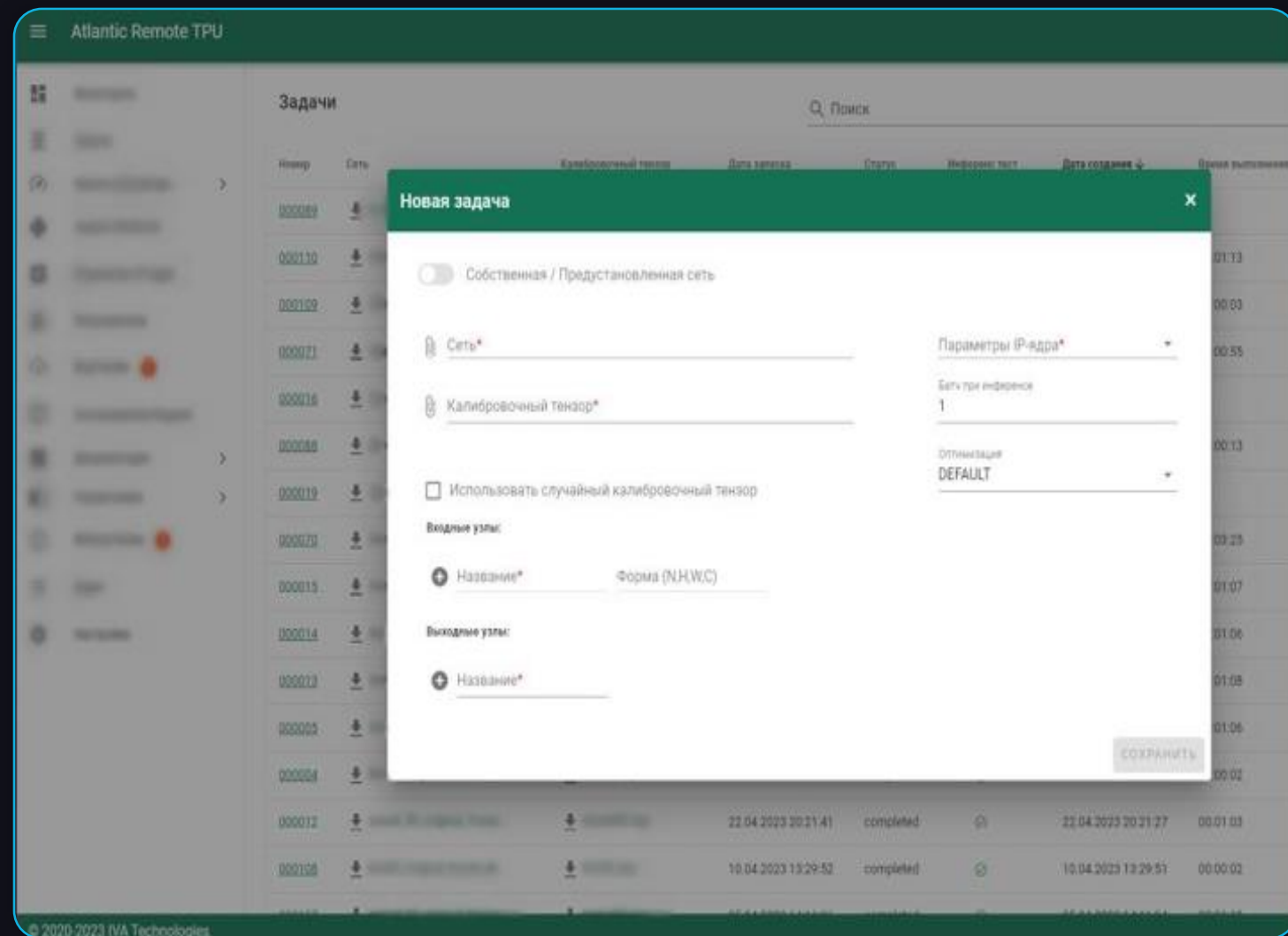
- ✓ Квантование модели ИНС
- ✓ Проверка точности квантованной модели
- ✓ Компиляция модели ИНС для TPU

## ИНФЕРЕНС:

- ✓ Эмуляция пре/пост-процессинга данных (Jupyter Notebook)
- ✓ Инференс ИНС на TPU

## АНАЛИТИКА:

- ✓ Снятие метрик производительности и точности инференса на TPU
- ✓ Проверка совместимости программных стеков





БЕСПИЛОТНЫЙ  
ТРАНСПОРТ



ИНТЕЛЛЕКТУАЛЬНАЯ  
ДОРОЖНАЯ  
ИНФРАСТРУКТУРА



ПОМОЩЬ В ПРИНЯТИИ  
РЕШЕНИЙ, ЦОДЫ ОБЩЕГО  
И СПЕЦИАЛЬНОГО  
НАЗНАЧЕНИЯ



ИНТЕЛЛЕКТУАЛЬНАЯ  
ВИДЕОАНАЛИТИКА



БИОМЕТРИЯ



ИНФОРМАЦИОННАЯ  
БЕЗОПАСНОСТЬ



УПРАВЛЕНИЕ БПЛА,  
ГРУППАМИ ОБЪЕКТОВ  
И Т.П.



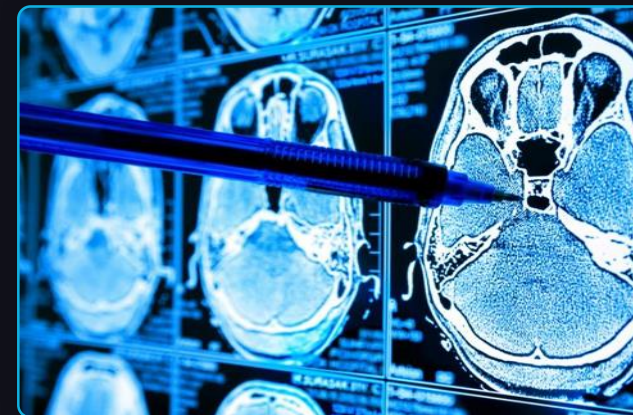
УМНЫЙ ГОРОД



Обеспечение принятия решений в центрах управления общего и специального назначения



Информационная безопасность, доверенные ПАК КИИ



Обработка медицинских данных, помощь врачу в выявлении и прогнозировании болезней



Контроль доступа, предупреждение о запрещенных действиях, нахождение в запрещенных пространствах



Городские биометрические системы



Повышение эффективности досмотровых систем безопасности



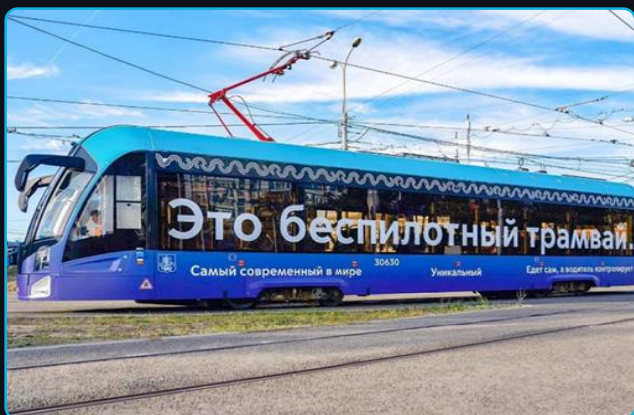
Контроль соблюдения ПДД



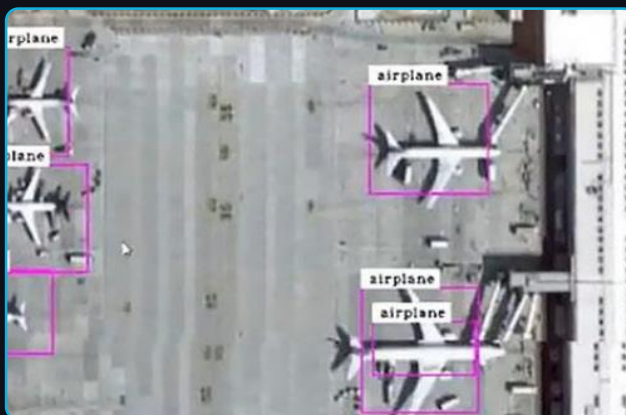
Система Безопасный город



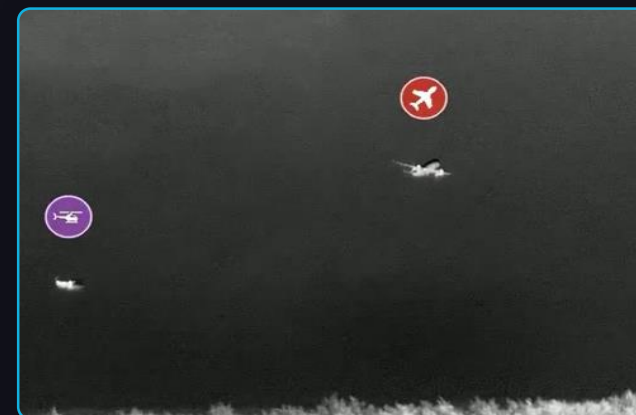
Метро без машиниста



Беспилотный наземный транспорт



Мониторинг обстановки и угроз, распознавание  
целей, автоматическое нанесение объектов  
на карты и передача в АСУ



Оптическое выявление и классификация  
слаборазличимых целей



Автономное и полуавтоматическое управление транспортом, БПЛА и роями дронов



Управление БПЛА в условиях сильных помех и в автономном режиме

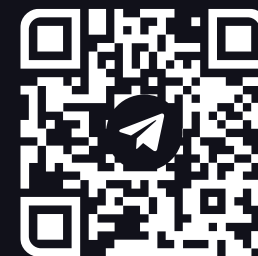


ФЕДЕРАЛЬНЫЙ ЦЕНТР  
ПРИКЛАДНОГО РАЗВИТИЯ  
ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

# СПАСИБО ЗА ВНИМАНИЕ!



ФЦПРИИ.РФ



t.me/fcprii