

Импульсные нейронные сети и нейропроцессоры

Михаил Киселев

Руководитель Лаборатории нейроморфных вычислений ЧГУ

Научный руководитель Лаборатории нейроморфных систем
искусственного интеллекта (Цифрум, ЧГУ, Мотив-ИТ, Kaspersky)



Бионический принцип в искусственном интеллекте

История

1. Нейронные сети первого поколения. Персептрон. Первоначальная идея состояла в имитации биологических нейронов и нейронных ансамблей, но получилось совершенно непохоже.
2. Экспертные системы. Полный отказ от бионического принципа. Идея строить интеллектуальные системы на основе формального логического вывода и методов индукции (потом добавился еще data mining).
3. Нейронные сети второго поколения. Глубокое обучение и сверточные сети. Вопрос о близости этих моделей к биологической реальности не имеет существенного значения.
4. Импульсные нейронные сети. Биологическая правдоподобность как один из основных принципов.

Строить искусственный интеллект, моделируя «элементную базу» естественного интеллекта.

Почему это снова актуально?

Победный марш нейронных сетей и их проникновение во все сферы жизни продолжается. Одновременно становятся ясны присущие им проблемы и ограничения. Эти проблемы в первую очередь проявляются на 2 противоположных полюсах искусственного интеллекта, где как раз ожидается технологический прорыв от его применения:

- Большие интеллектуальные системы по когнитивным возможностям приближающиеся к человеческим. «Большой» интеллект.
- Миниатюрные автономные интеллектуальные устройства для применения во всех сферах жизни (интернет вещей, робототехника, безопасность, медицинские устройства, встраиваемые в тело человека и т.д.). «Малый» интеллект.

Эти проблемы можно разбить на три группы, хотя они в большой степени взаимосвязаны:

- Энергоэкономичность
- Масштабируемость
- Необходимость взаимодействовать с динамической асинхронной средой.

Энергоэкономичность

Трудно в точности сравнить затраты энергии на проведение одних и тех же операций биологического мозга и компьютерных реализаций традиционных сетей, но очевидно, что мозг на много порядков экономичнее. Мозг потребляет 20-30 Вт энергии, в то же время все GPU суперкомпьютера «Кристафари», которому потребовалось несколько месяцев для обучения нейросетевой лингвистической модели GPT-3, потребляют около 1 МВт.

Это в равной степени препятствует расширению применения традиционных нейросетей и в «малом» и в «большом» интеллекте.

Масштабируемость

Очевидно, что решение более сложных интеллектуальных задач неизбежно сопровождается увеличением сложности и размера решающих их нейросетей. Если принять за единицу сложности одну синаптическую связь в мозге и один параметр нейросетевой модели, то при всей приблизительности этой аналогии ясно, что для достижения сравнимых с человеком когнитивных возможностей требуется увеличение роста сложности нейросетевых моделей на несколько порядков.

Число синаптических связей в мозге человека $\sim 10^{14}$.

Число параметров в самых сложных современных нейросетевых моделях $\sim 10^{12}$. Обучение такой модели требует месяцев счета на самых мощных имеющихся суперкомпьютерах. Дальнейшее увеличение сложности нейросетевых моделей традиционного типа и на существующих аппаратных платформах даже на порядок представляется проблематичной.

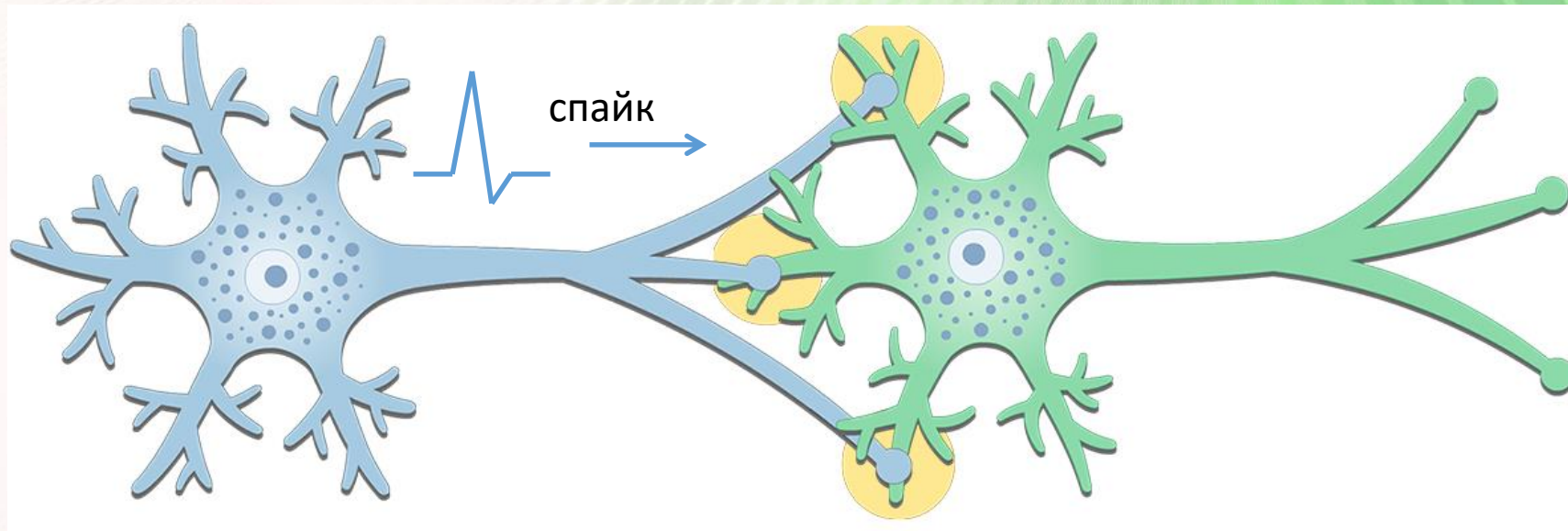
Необходимость взаимодействовать с динамической асинхронной средой

Традиционные нейронные сети предполагают очень жесткий протокол взаимодействия с их информационным окружением. Предполагается, что входные данные представляют собой последовательность фреймов данных. После предъявления очередного фрейма состояние всей нейросети синхронным образом перевычисляется и сеть возвращает некоторый выходной набор значений. В то же время человек и любые интеллектуальные агенты, взаимодействующие с реальным миром, должны жить в непрерывном времени в условиях динамических потоков данных, разворачивающихся в разных временных шкалах.

Импульсные нейронные сети (ИНС) – класс формальных моделей нейронных ансамблей мозга, позволяющих решить перечисленные проблемы с помощью тех же принципов, что лежат в основе функционирования биологических нейронов.

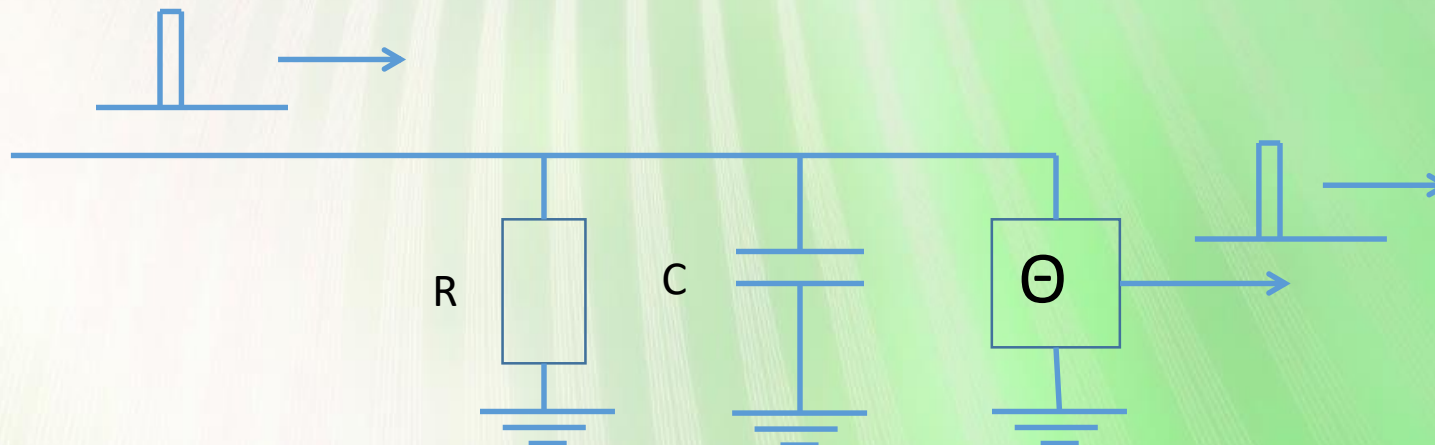
- Информация, которая передается от нейрона к нейрону – это объект, не имеющий никаких других атрибутов кроме времени его генерации. В мозге ему соответствует нервный импульс или спайк, передаваемый от нейрона к нейрону через синаптическую связь. Функционирование нейрона проявляется только как генерация им спайков. Время генерации нейроном спайков зависит только от спайков, пришедших на его синапсы.
- Функционирование разных нейронов никак явно между собой не синхронизировано.
- **Энергоэкономичность** – основанная на событиях модель функционирования нейрона – функционирование нейрона состоит в выполнении им простых операций после прихода пресинаптического спайка, после которых он переходит в пассивное состояние не требующее вычислений и не расходующее энергию. Простота объектов, используемых для передачи информации между нейронами (элементарные события – спайки), определяет простоту и быстроту операций их обработки.
- **Масштабируемость** – отсутствие явной синхронизации работы нейронов, отсутствие постоянного потока данных по всем синаптическим связям, толерантность к потере отдельных спайков и отказу отдельных нейронов
- **Асинхронная среда** – нет принципиальной разницы между взаимодействием между нейронами и их взаимодействием с окружающей средой. В обоих случаях это асинхронный обмен спайками. Протокол этого обмена никак не фиксируется.

Простейшая формальная модель импульсного нейрона: нейрон – пороговый интегратор с утечкой (LIF)



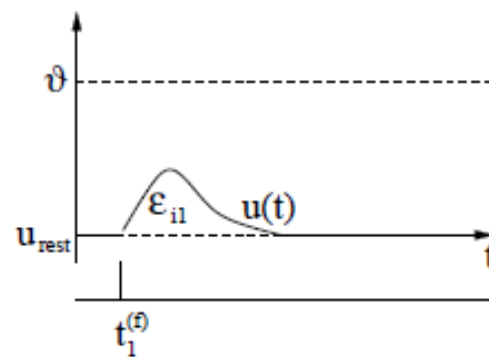
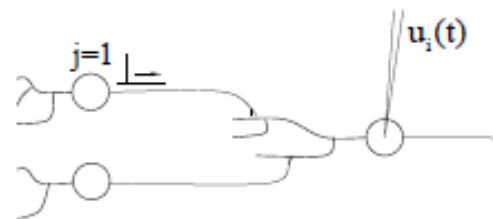
Самая базовая операция, на которой построено функционирование ИНС, – фиксация совпадений прихода спайков на разные синапсы.

Уже существуют **нейрочипы**, которые реализуют эту модель «в железе»

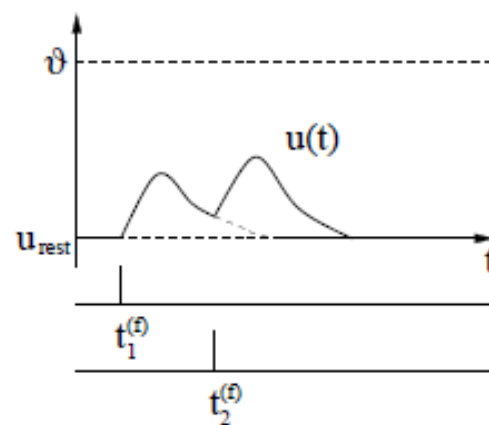
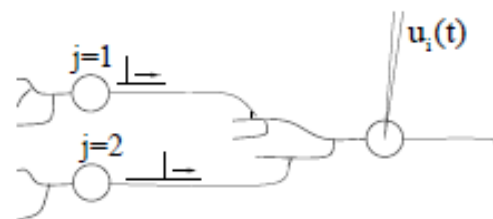


Например, отечественный нейрочип **«Алтай»**, разработанный компанией Мотив НТ

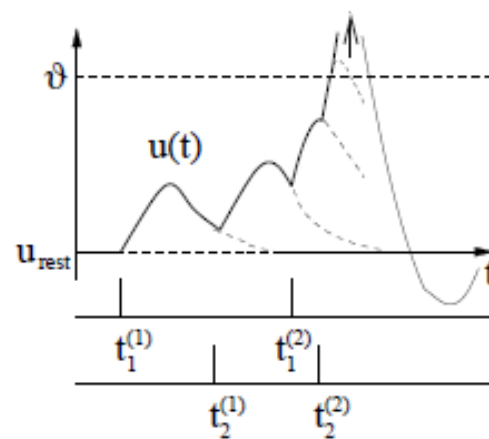
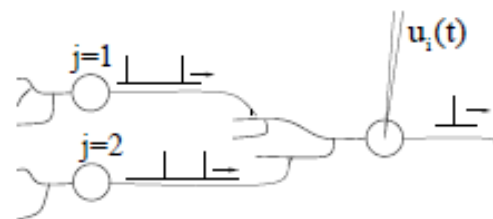
A



B



C



Генерация спайка

Класс моделей SRM₀

1. «Одномерная» модель – состояние нейрона описывается его мембранным потенциалом.
2. Мембранный потенциал зависит только от времени прихода спайков и времени генерации последнего спайка.
3. Вес синапса определяет величину инжектируемого заряда.

$$u_i(t) = \eta(t - \hat{t}_i) + \sum_j \sum_f \epsilon_{ij}(t - t_j^{(f)}) + u_{\text{rest}}$$

$$u_i(t) = \vartheta \text{ and } \frac{d}{dt}u_i(t) > 0 \quad \Longrightarrow \quad t = t_i^{(f)}$$

Нейроны с синапсами, контролирующими мембранную проводимость

$$\begin{cases} \frac{du}{dt} = -cu - c_I(u - u_I) + c_E(u_E - u) \\ \frac{dc_I}{dt} = -\frac{c_I}{\tau_I} + \sum_i w_i^- \sum_f \delta(t - t_i^{(f)}) \\ \frac{dc_E}{dt} = -\frac{c_E}{\tau_E} + \sum_i w_i^+ \sum_f \delta(t - t_i^{(f)}) \end{cases}$$

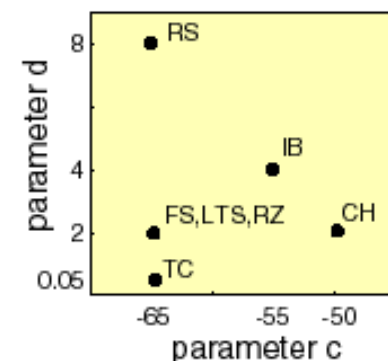
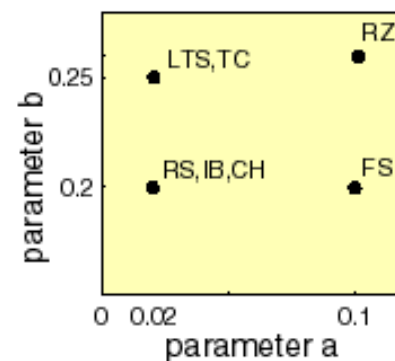
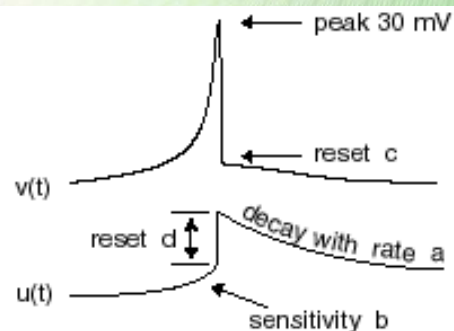
Если $u = u_{THR}$, генерация спайка, $u \leftarrow u_0$

Нейрон Ижикевича

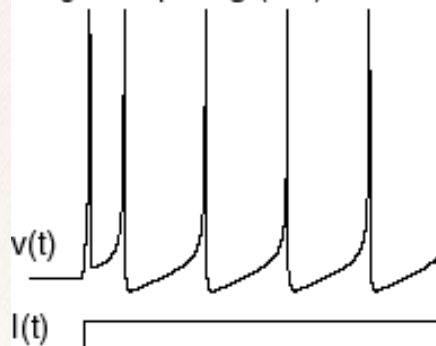
$$v' = 0.04v^2 + 5v + 140 - u + I$$

$$u' = a(bv - u)$$

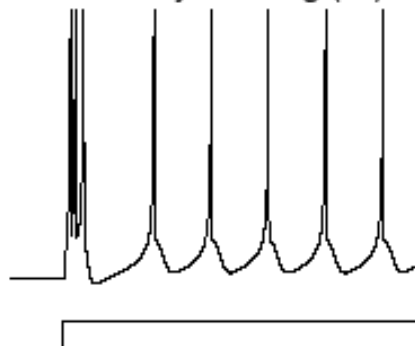
if $v = 30$ mV,
then $v \leftarrow c$, $u \leftarrow u + d$



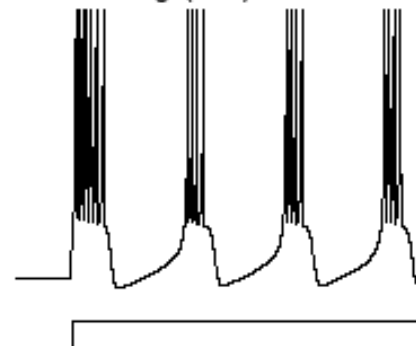
regular spiking (RS)



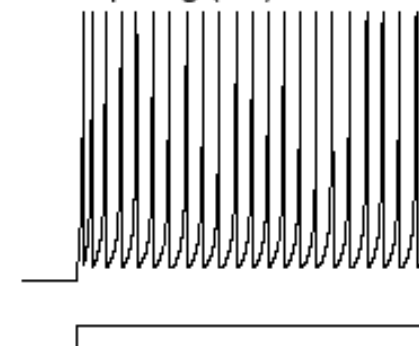
intrinsically bursting (IB)



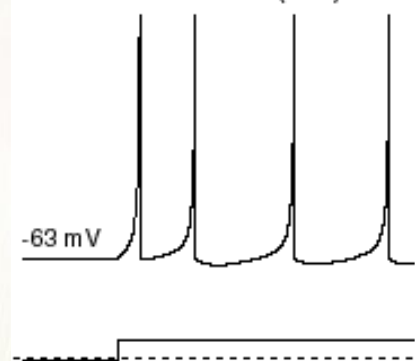
chattering (CH)



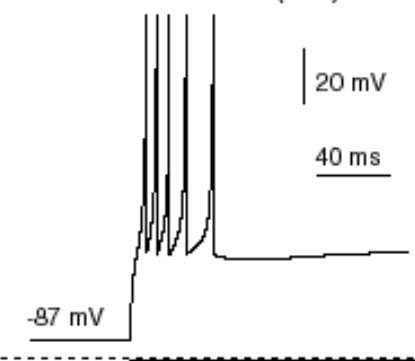
fast spiking (FS)



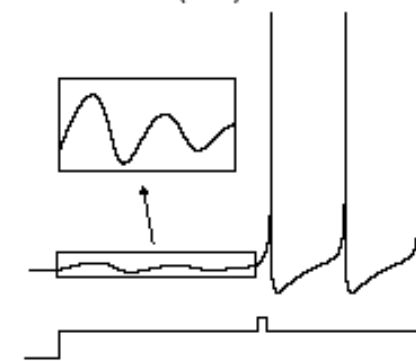
thalamo-cortical (TC)



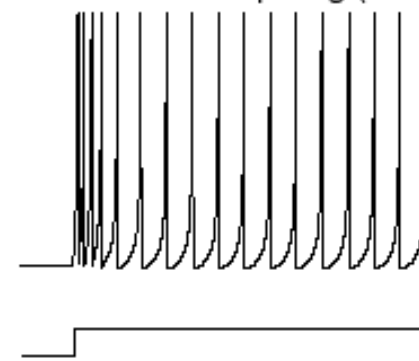
thalamo-cortical (TC)



resonator (RZ)



low-threshold spiking (LTS)



Детальная модель нейрона Hodgkin, Huxley

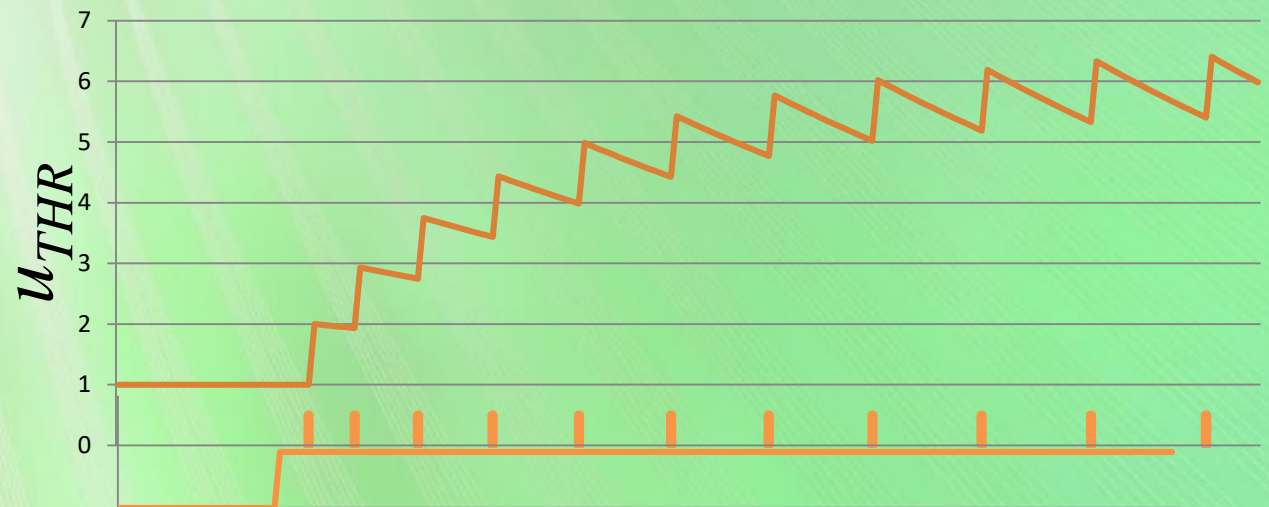
$$C \frac{dv}{dt} = -\underline{g}_{Na} m^3 h (v - V_{Na}) - \underline{g}_K n^4 (v - V_K) - g_r (v - V_r) + I_{ap} \quad \frac{dn}{dt} = \alpha_n(v)(1 - n) - \beta_n(v)n$$
$$\frac{dm}{dt} = \alpha_m(v)(1 - m) - \beta_m(v)m \quad \frac{dh}{dt} = \alpha_h(v)(1 - h) - \beta_h(v)h$$

Пороговый интегратор с утечкой и адаптивным порогом

$$\begin{cases} \frac{du}{dt} = -cu + \sum_i w_i \sum_f \delta(t - t_i^{(f)}) \\ \frac{du_{THR}}{dt} = -\frac{u_{THR} - \bar{u}_{THR}}{\tau_T} + \sum_k \hat{T} \delta(t - \hat{t}_k) \end{cases}$$

- Если $u(t) = u_{THR}(t)$, генерация спайка, $u \leftarrow u_0$
- $u(t)$ ограничен снизу значением u_L

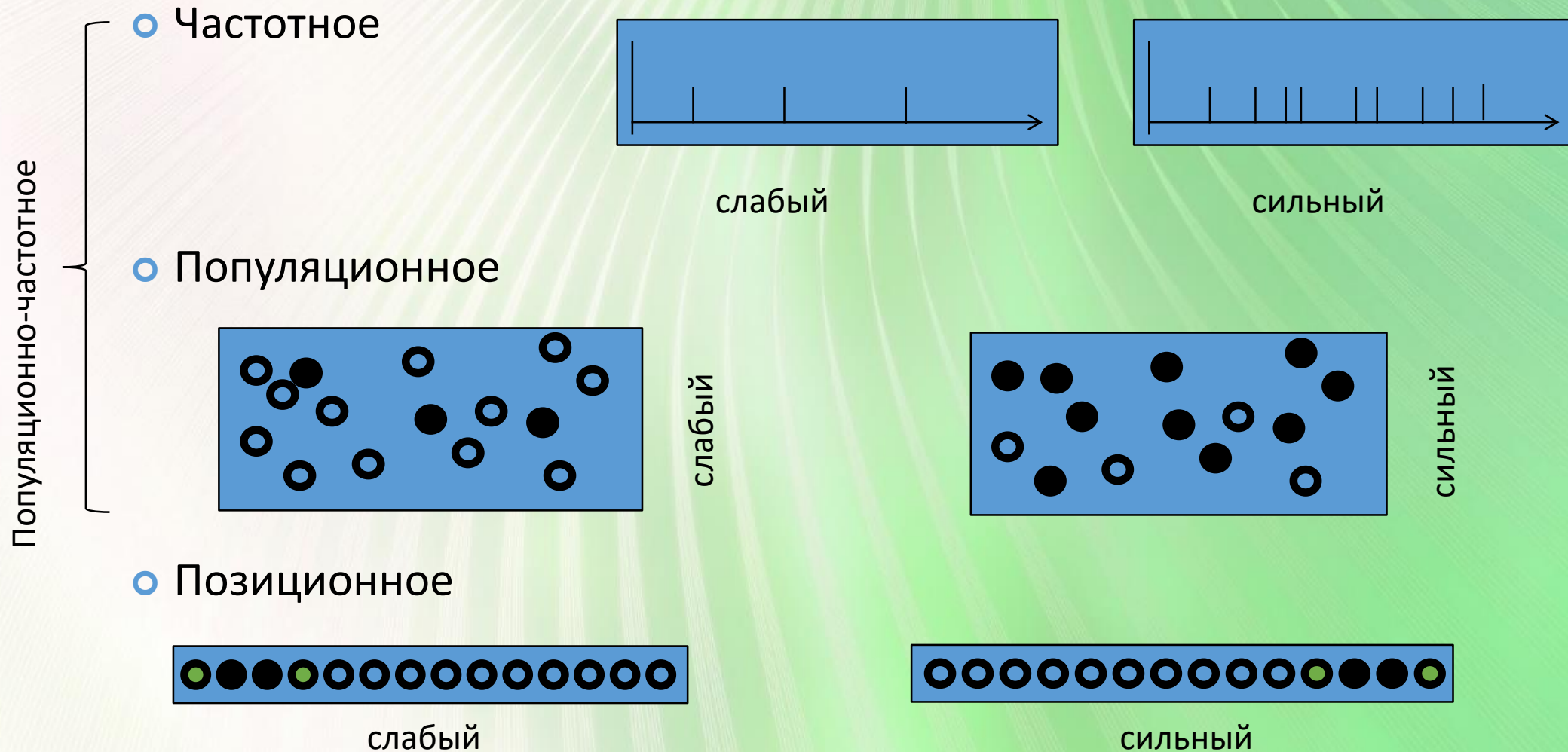
- Гомеостаз – более слабый рост частоты генерации спайков при растущей стимуляции
- Более выраженная реакция на динамические стимулы



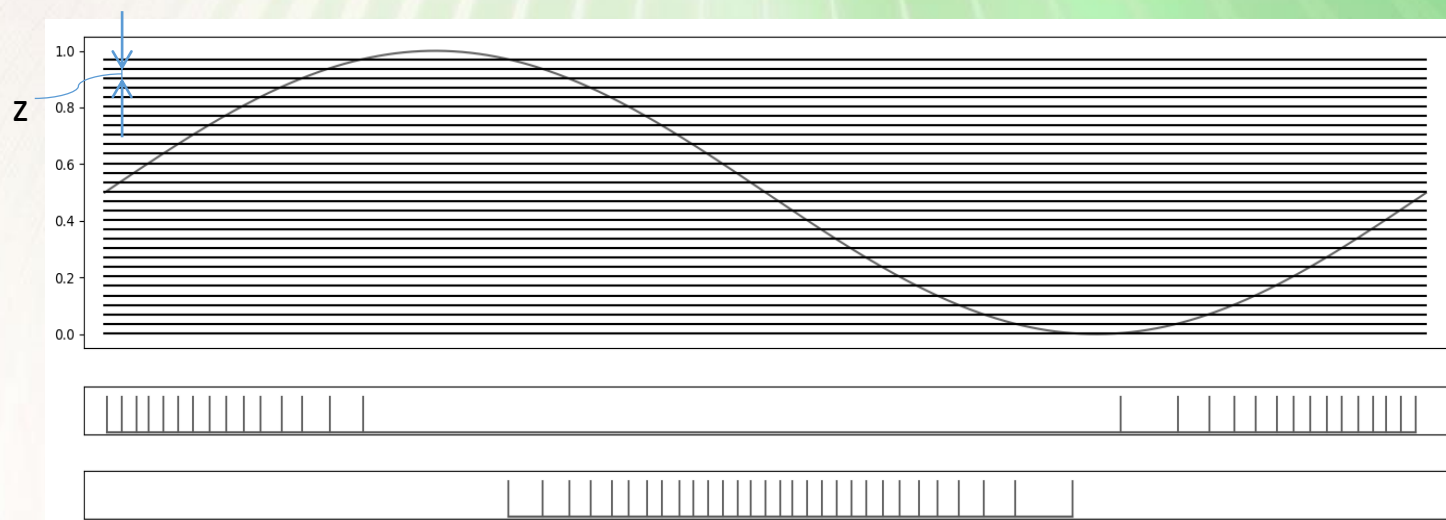
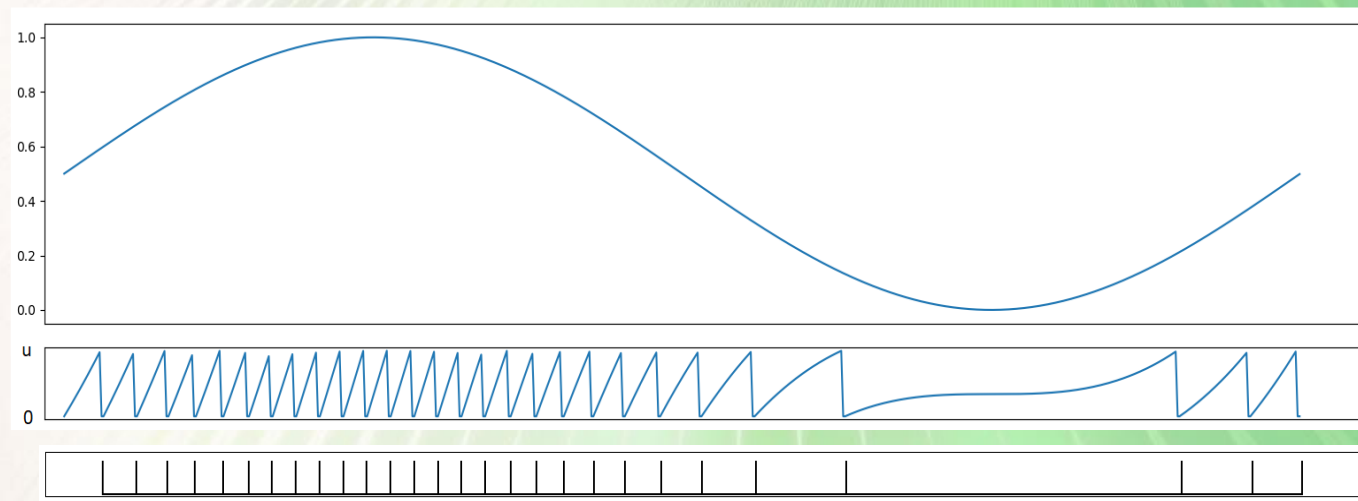
Модели нейронов с точки зрения аппаратной реализации

- Аналоговая реализация экспонент, аддитивных компонент уравнений, сравнения с пороговым потенциалом, переустановки мембранного потенциала
- Цифровая реализация — малое количество мультипликативных операций, отсутствие трансцендентных функций
- Линеаризованные модели нейронов:
 - замена экспонент линейным ростом (падением)
 - использование аппаратно сгенерированных случайных чисел

Асинхронные методы кодирования информации в ИНС



Дифференциальное кодирование



Синхронные методы кодирования информации в ИНС

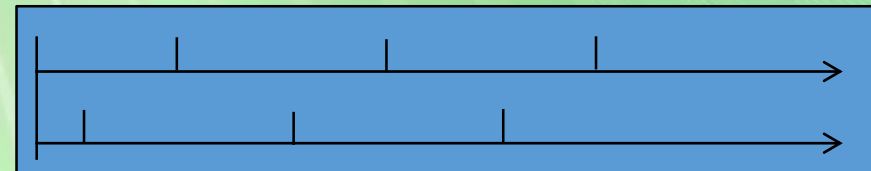
- Пространственно-временное



- Временной сдвиг

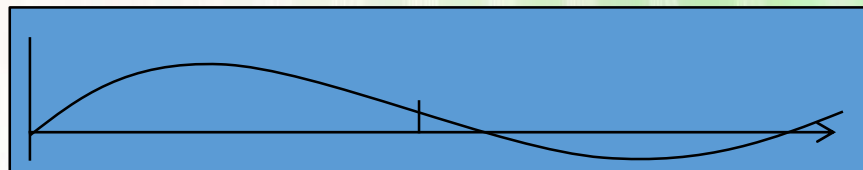


слабый

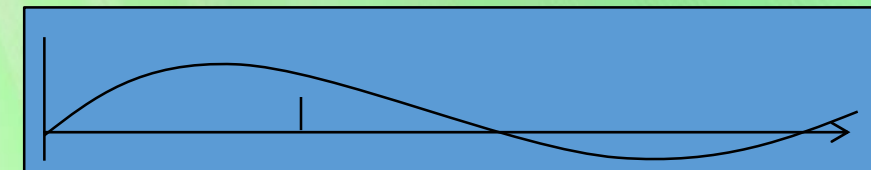


сильный

- Фазовое



слабый



сильный

Архитектуры ИНС

большое разнообразие

- Хаотические рекуррентные ИНС
- Слоистые с латеральным торможением («победитель получает все»)
- Модели слоисто-колончатых нейронных структур коры головного мозга
- ...

Обучение ИНС

общая идея такая же, как для традиционных сетей – модификация синаптических весов, но:

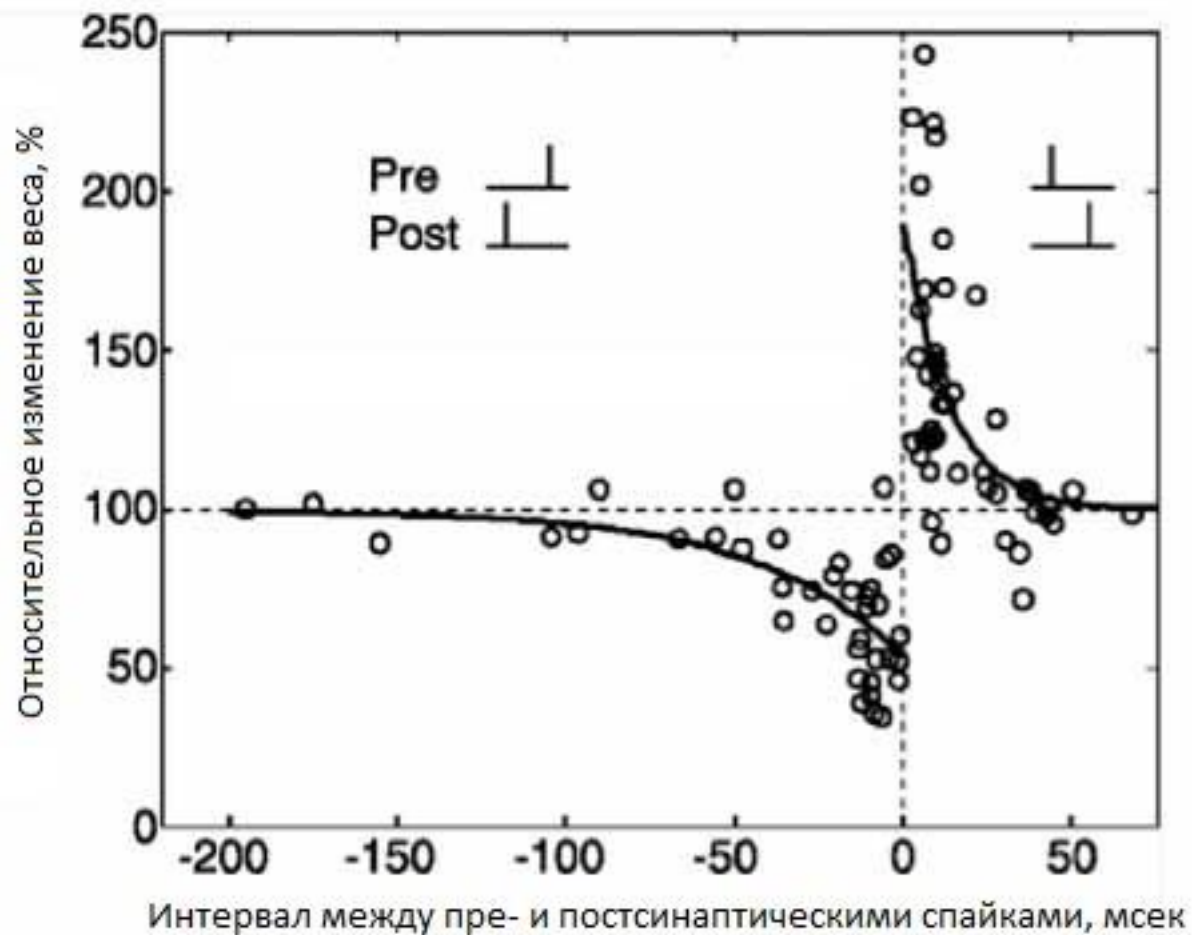
- Дискретное функционирование препятствует непосредственному применению градиентных методов (аналогов алгоритма обратного распространения ошибки). Хотя проекции основных подходов обучения традиционных нейросетей на ИНС существуют и развиваются.
- Коллективное кодирование информации и принцип толерантности к отказам отдельных нейронов делает практически ненаблюдаемым эффект модификации единичного синаптического веса на глобальные показатели поведения сети.
- Алгоритм обратного распространения ошибки биологически нереалистичен.

Основные принципы обучения ИНС:

Применяются еще методы с динамически меняющейся структурой сети (добавление нейронов, которые дали бы правильный ответ и уничтожение нейронов, которые часто ошибаются)

- Принцип локальности
- Принцип Хебба

STDP – базовая модель синаптической пластичности



$$\Delta w = \begin{cases} A_+ e^{-\frac{\Delta t}{\tau_+}}, & \text{если } \Delta t > 0 \\ -A_- e^{\frac{\Delta t}{\tau_-}}, & \text{если } \Delta t < 0 \end{cases}$$

По Bi G., Poo M. 1998

Гомеостатические поправки к STDP

- Стабилизация синапса за счет введения понятия синаптического ресурса

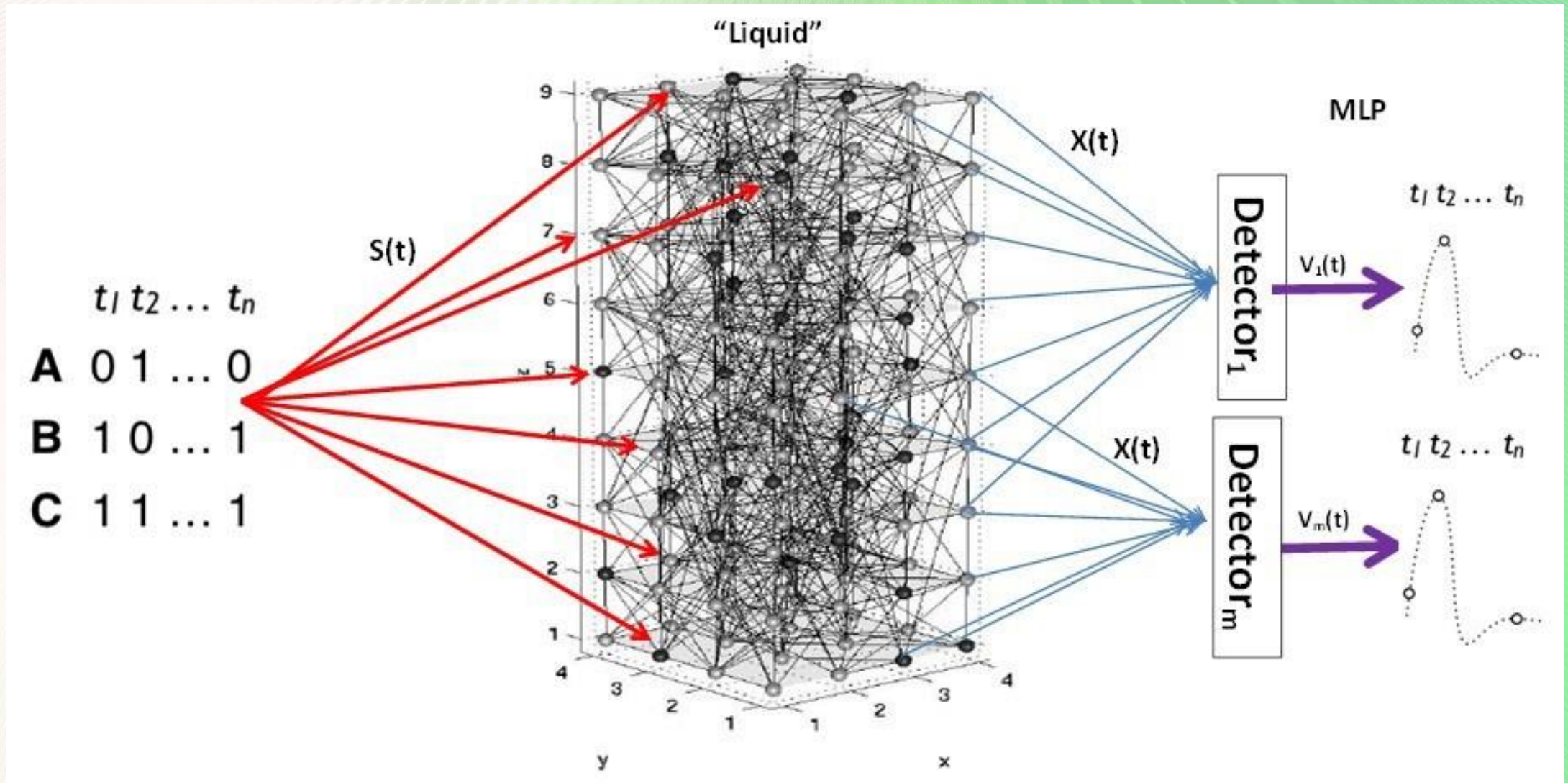
$$w = w_{\min} + \frac{(w_{\max} - w_{\min}) \max(W, 0)}{w_{\max} - w_{\min} + \max(W, 0)}$$

- ограничение суммарного синаптического веса (или константный суммарный синаптический вес)
- уменьшение эффективности **LTP** при повышении частоты генерации спайков
- модуляция пластичности активностью определенных синапсов
- Асимметричная **STDP** (безусловное подавление синапса при приходе спайка)
- моделирование кратковременного синаптического подавления

Общий подход к обучению ИНС – обучение, основанное на выявлении корреляций

- **Unsupervised learning** – поиск корреляций во входном сигнале
- **Supervised learning** – поиск корреляций между входным сигналом и сигналом, кодирующим целевую переменную
- **Reinforcement learning** – поиск корреляций между входным сигналом, выходным сигналом и оценочным сигналом (возможно – с временной задержкой)
- Анतिकорреляции находятся аналогичным образом за счет использования тормозных нейронов
- Обучающиеся структуры могут иерархическими – например, **supervised learning** может быть организовано «поверх» структур, реализующих выделение информативных свойств с помощью **unsupervised learning**

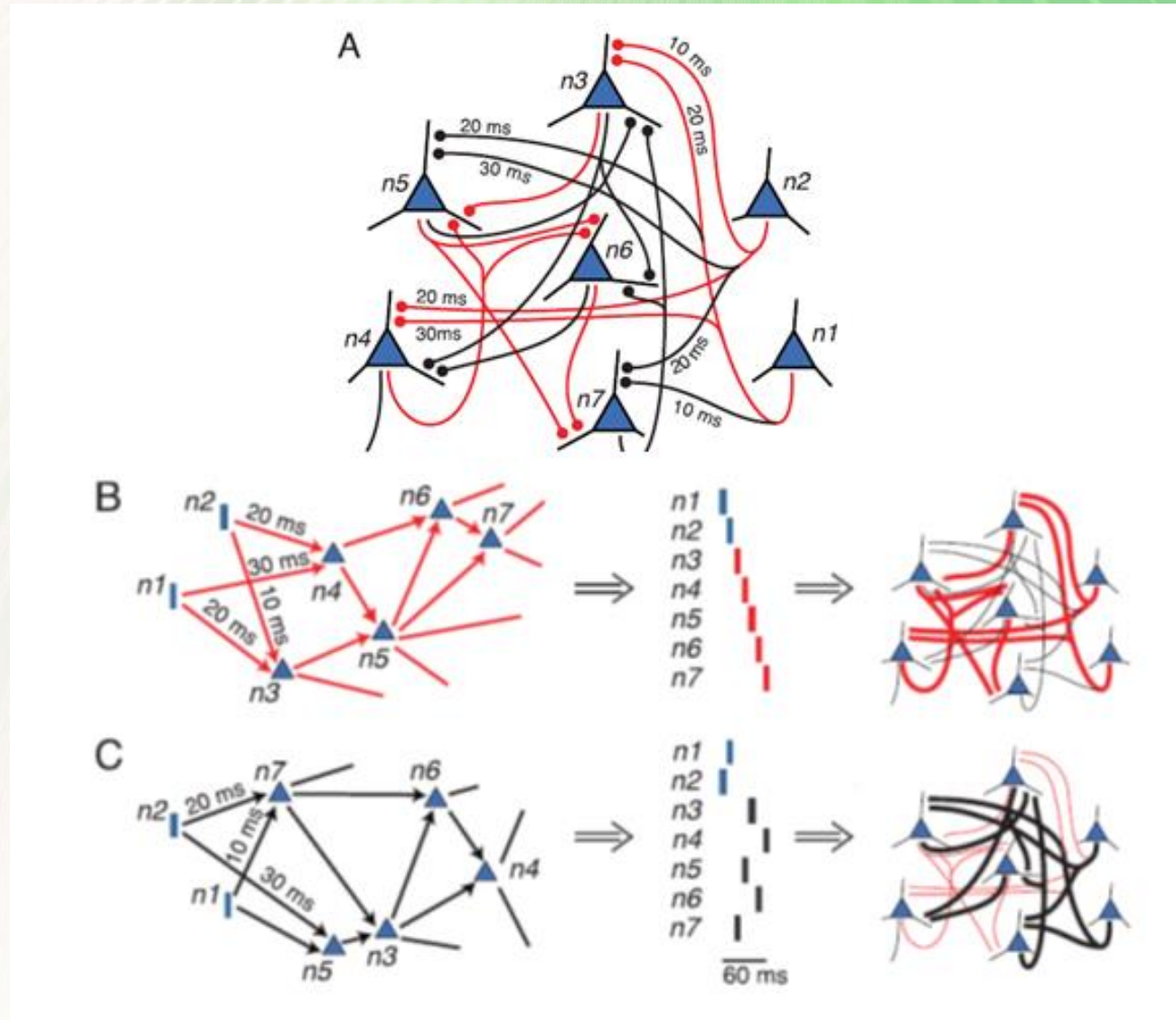
Альтернативный подход к обучению – машина с жидким состоянием



Реализация оперативной памяти в ИНС

1. Рекуррентные структуры с циклической активностью
 1. Хорошо управляемая память
 2. Легкая считываемость
 3. Низкая емкость
 4. Может запоминать небинарные величины
2. «Медленные» компоненты состояния нейронов
 1. Не очень надежная память
 2. Непростые механизмы считывания
 3. Сильно ограничена по времени
 4. Средняя емкость
 5. Запоминает лишь бинарные величины, но может хранить время
3. Полихронные группы и кратковременная пластичность
 1. Трудность запоминания
 2. Эффективно реализуется только в больших ИНС
 3. Очень большая емкость
 4. Запоминает лишь бинарные величины
 5. Сильно ограничена по времени
4. Непрерывные аттракторы
 1. Может хранить действительные числа
 2. Очень низкая емкость
 3. Легкая считываемость
 4. Большое время

Полихронные нейронные группы (по Szatmary B., Izhikevich E. Spike-Timing Theory of Working Memory, 2010)



Нейроморфные технологии и исследование мозга – связанные и взаимообогащающие научные направления



*

Для этого требуются нейрокомпьютеры. Пока они малодоступны – GPU-кластеры.

Аппаратные платформы для моделирования ИНС

Архитектура	Платформа	Кем реализуется
Цифровая, универсальный процессор	GPU, GPU-кластеры	ранообразные группы и проекты
	SpiNNaker	Университет Манчестера
Цифровая	Loihi	Intel
	TrueNorth	IBM
	Алтай	Мотив
	Tianji	Университет Синьхуа
Гибридная	NeuroGrid	Stanford
	BrainScaleS	НБР, Гейдельбергский университет

Моделирование ИНС на GPU

3 основных процесса, которые требуется моделировать

- обновление состояния каждого нейрона и определение факта генерации им спайка
- распространение спайков от пресинаптических нейронов к постсинаптическим нейронам
- модификация весов синапсов в соответствии с законами пластичности

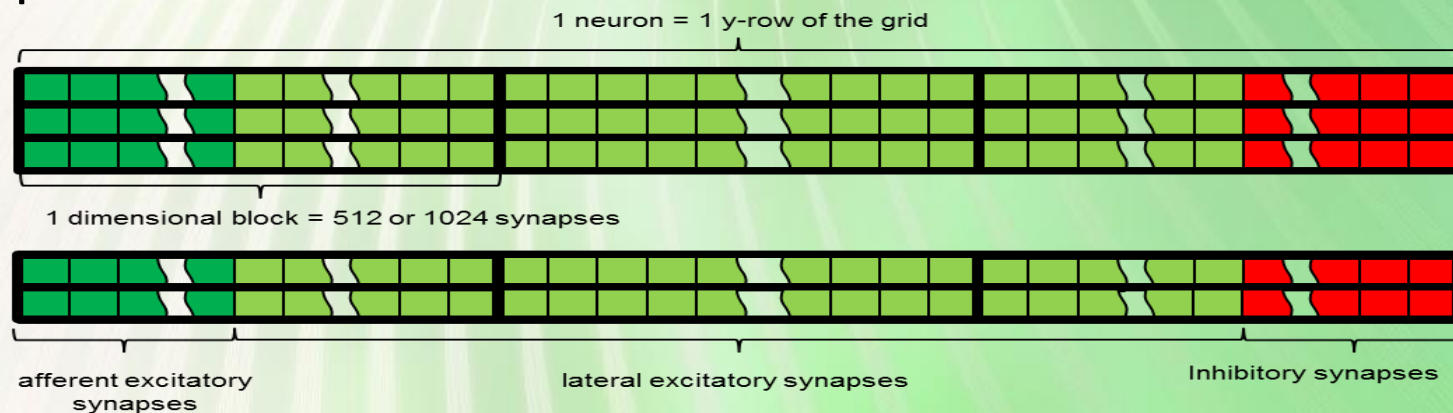
GPU с точки зрения моделирования ИНС

1. Физический уровень параллелизма – несколько тысяч ядер
2. Логический параллелизм обеспечивается иерархией структур потоков управления «нить – блок – решетка», абстрагированной от физической конфигурации вычислителя
3. Ограничение **1024** нити на блок влечет «мягкое» ограничение **1024** синапсов на нейрон (наиболее эффективная реализация)
4. Низкая цена создания и оперирования с нитью – простые вычисления на многочисленных синапсах
5. Регистровая память достаточна для хранения состояния нейронов в любых практически используемых моделях
6. Память на плате – главный лимитирующий фактор (**~10 GB**), достаточна для хранения очередей пресинаптических спайков и данных по задержкам и весам сотен миллионов синапсов
7. Возможны трудности с подачей интенсивного потока внешних спайков в силу медленности и высокой латентности **PCI-E**

Вывод: GPU – адекватное средство для моделирования ИНС размером до 100000 нейронов, 10000 синапсов на нейрон

Принципы реализации ИНС на GPU

1. Необходимость частичной синхронизации – каждая фаза пересчета сети как отдельный вызов **kernel**.
2. Наиболее удобный подход – 1 нейрон \leftrightarrow 1 блок, 1 синапс \leftrightarrow 1 нить. Менее эффективно (при большем числе синапсов) – 2-мерная решетка, 1 нейрон \leftrightarrow 1 Y-слой решетки.

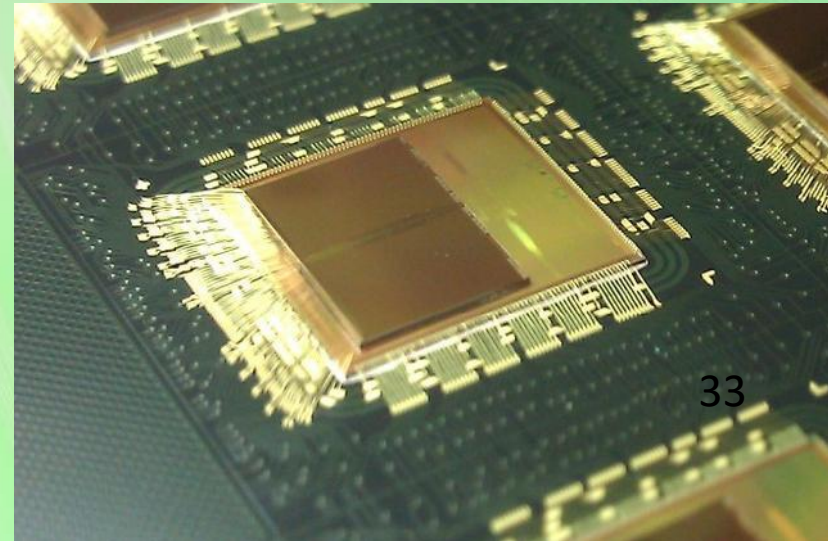
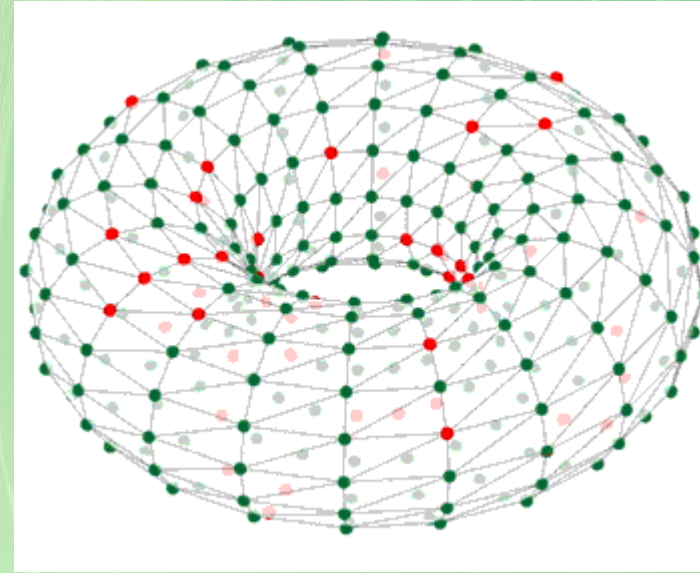
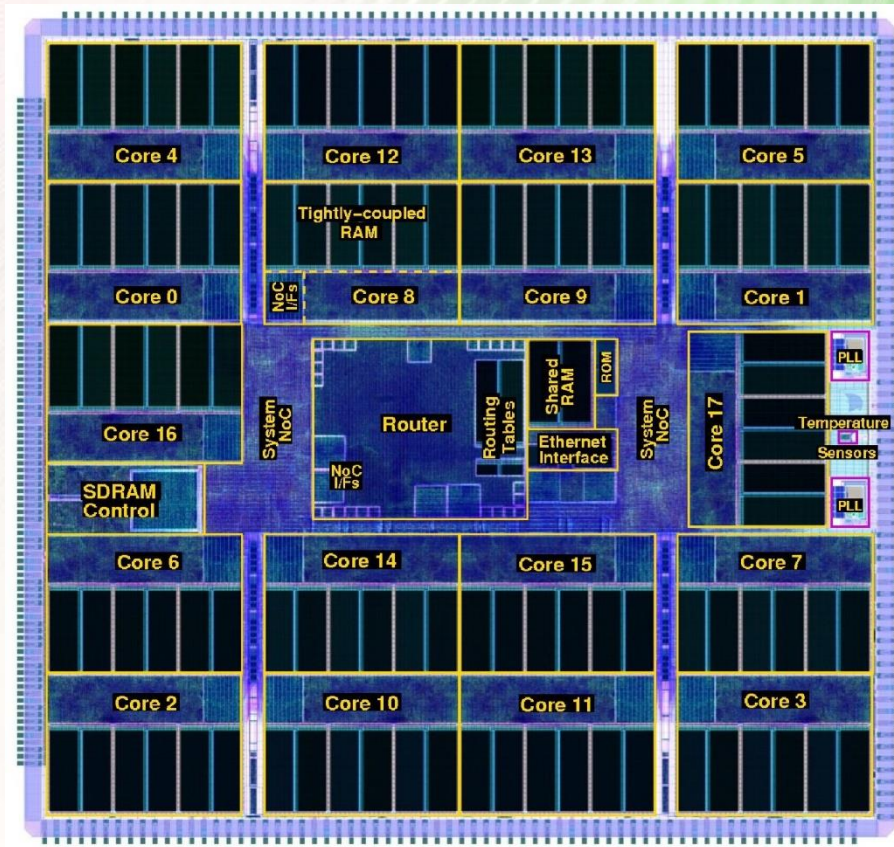


3. Эмуляция задержек распространения спайков за счет входных битовых очередей (задержки в интервале **1-64**)

Обмен спайками по протоколу AER

- Разумная альтернатива физическим связям
- Обычно крайне невелико информационное наполнение – обычно только идентификатор пресинаптического нейрона
- Доставка пакета не гарантируется
- Доставка пакета предполагается мгновенной, так что синаптические задержки моделируются на стороне постсинаптического нейрона

SpiNNaker



SpiNNaker – характеристики

- **18** асинхронно работающих **ARM968** ядер на чипе, из них **1** выполняет системные функции и **1** зарезервировано на случай сбоя ядра
- **128 MB SDRAM** смонтирована на втором слое чипа
- Эффективная схема маршрутизации **AER** пакетов
- **6** линий для связи с другими чипами
- Программная поддержка большинства используемых моделей нейронов, синапсов и синаптической пластичности
- **1** чип способен эмулировать в реальном времени несколько тысяч пластичных нейронов с сотнями синапсов
- Виртуальная конфигурация структуры сети
- Основной структурный элемент - **48**-чиповая плата

SpiNNaker – масштабируемость



48-чиповая плата, $\sim 10^5$ нейронов



стойка из 24 плат, $\sim 10^6$ нейронов
энергопотребление 2 kW



нейросуперкомпьютер SpiNNaker – 25 стоек, 518400 ядер с виртуальной тороидальной топологией, $\sim 10^8$ нейронов (планируется удвоить).

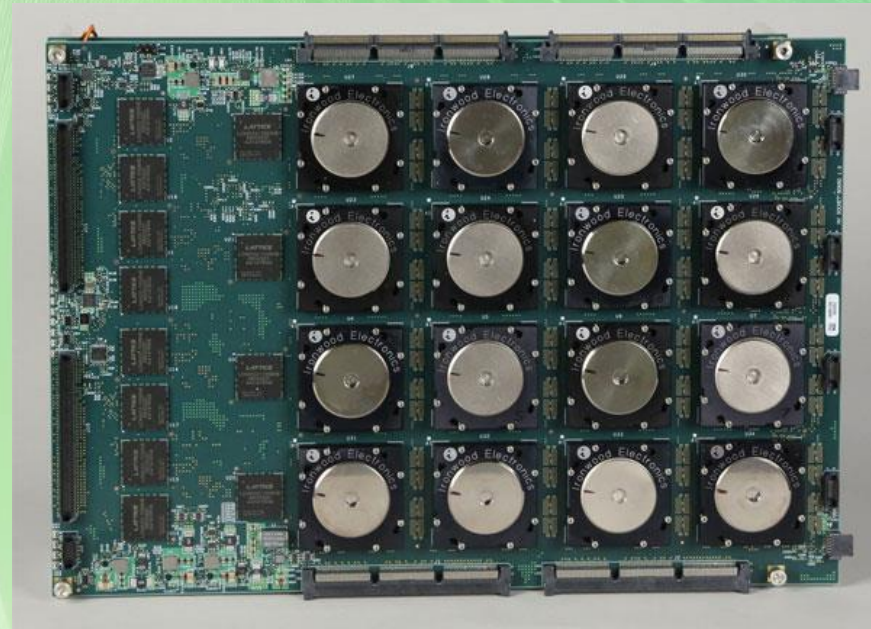
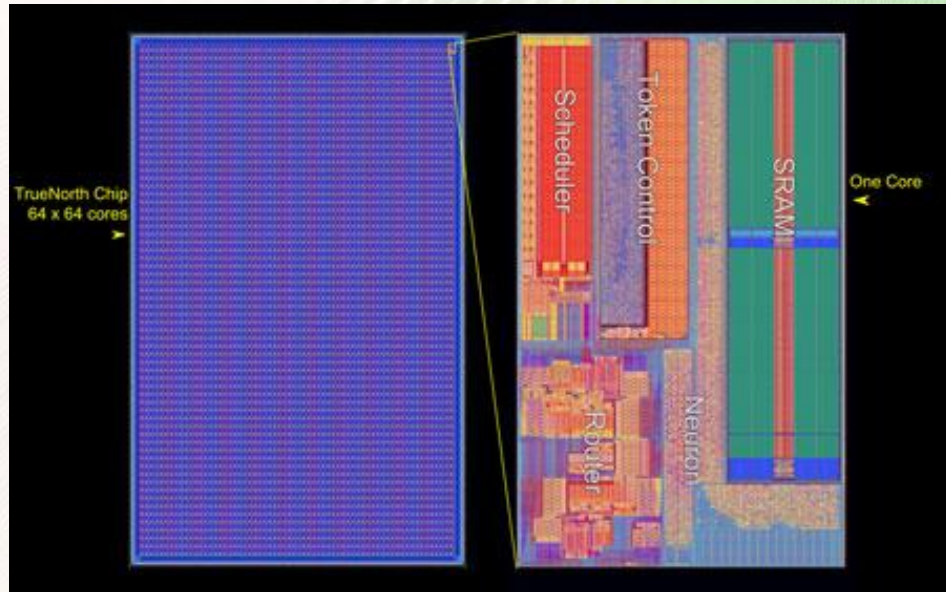
SpiNNaker – проблемы

- скалярный универсальный процессор => много неиспользуемых элементов => недостаточное количество эмулируемых нейронов на чипе
- доступ к внешней памяти – «узкое место» фон-Неймановской архитектуры
- недостаточная системная поддержка, трудность модификации стандартных моделей нейронов, отладки, мониторинга

Гибридные нейроморфные вычислители

- **Neurogrid – Stanford - 1M** нейронов,
6B синапсов
- **BrainScaleS – Heidelberg - 200k**
нейронов, **50M** синапсов

TrueNorth



TrueNorth – достигнутые параметры

- **1** миллион нейронов на чипе
- **256** синапсов на нейрон
- Модель нейрона – обобщенный линеаризованный стохастический **LIF**
- Самый крупный чип современности – **5.4** миллиардов транзисторов; **28nm** процесс
- **4096** параллельно работающих ядер
- **5MB** памяти на чипе
- **70 mW** – пиковая потребляемая мощность, **20 mW/cm²** – сравнимо с мозгом.

TrueNorth – проблемы

- Модель нейрона простая и жесткая (не программируется).
- Нет синаптической пластичности.
- Синапсов слишком мало.
- Архитектура сети жестко ориентирована на сверточные сети.
- Веса синапсов могут принимать лишь **1** из **4** значений.

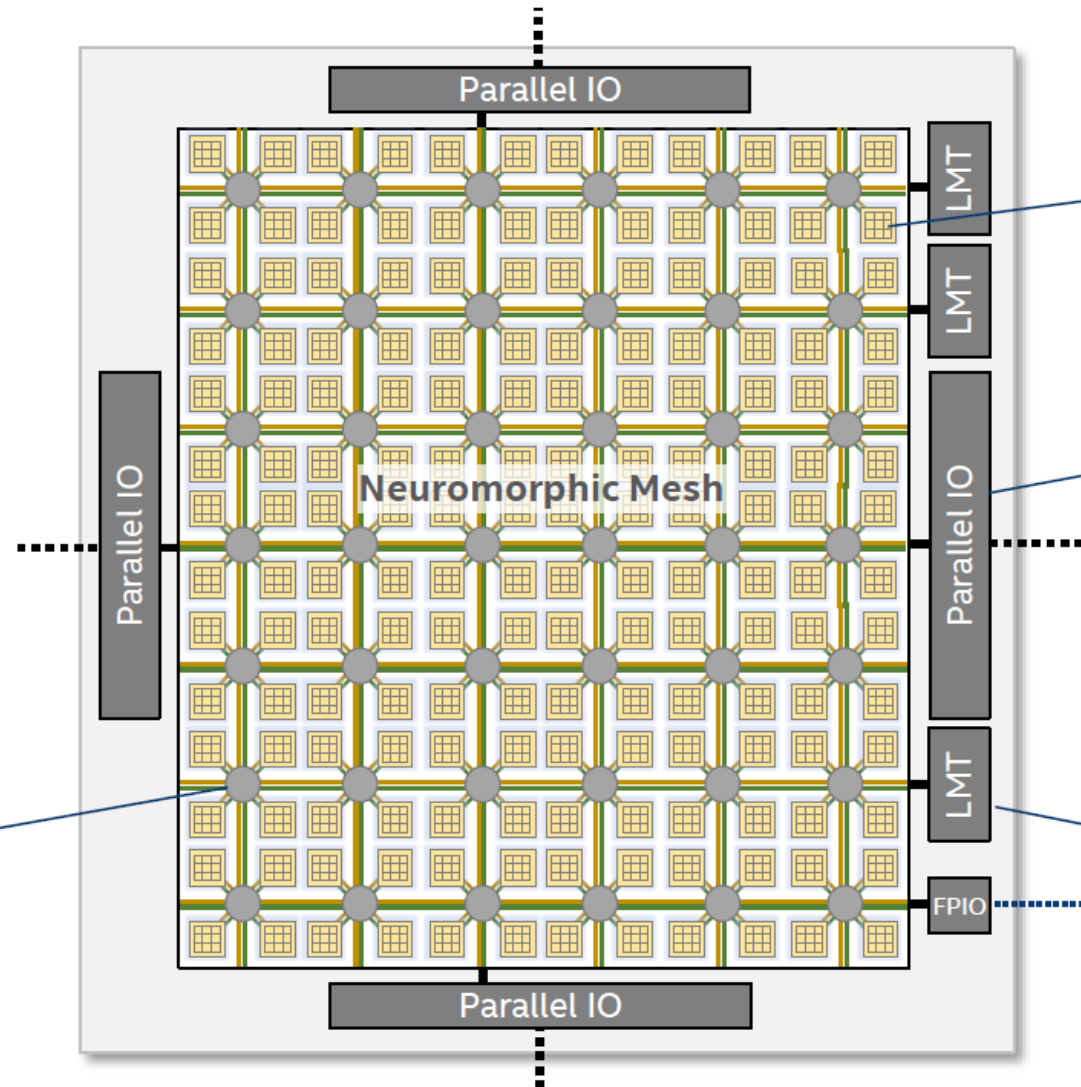
Loihi

- 130 тысяч нейронов с 130 млн. синапсов;
- 128 асинхронно асинхронно работающих нейроядер + 3 универсальных процессора Lakemont;
- техпроцесс 14-нм;
- в сравнении с GPU энергопотребление ниже в сотни раз на бенчмарках по deep learning;
- Реализована синаптическая пластичность на чипе для реализации локальных алгоритмов обучения.

Архитектура чипа

Technology:	14nm
Die Area:	60 mm ²
Core area:	0.41 mm ²
NmC cores:	128 cores
x86 cores:	3 LMT cores
Max # neurons:	128K neurons
Max # synapses:	128M synapses
Transistors:	2.07 billion

- Low-overhead NoC fabric**
- 8x16-core 2D mesh
 - Scalable to 1000's cores
 - Dimension order routed
 - Two physical fabrics
 - 8 GB/s per hop



Neuromorphic core

- LIF neuron model
- Programmable learning
- 128 KB synaptic memory
- Up to 1,024 neurons
- Asynchronous design

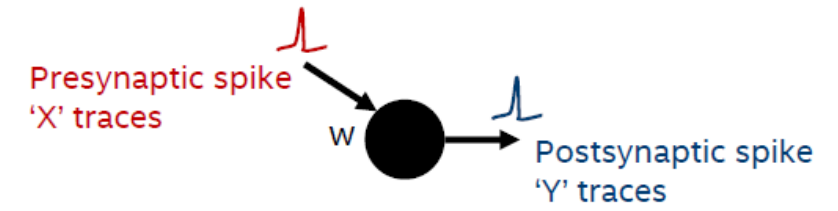
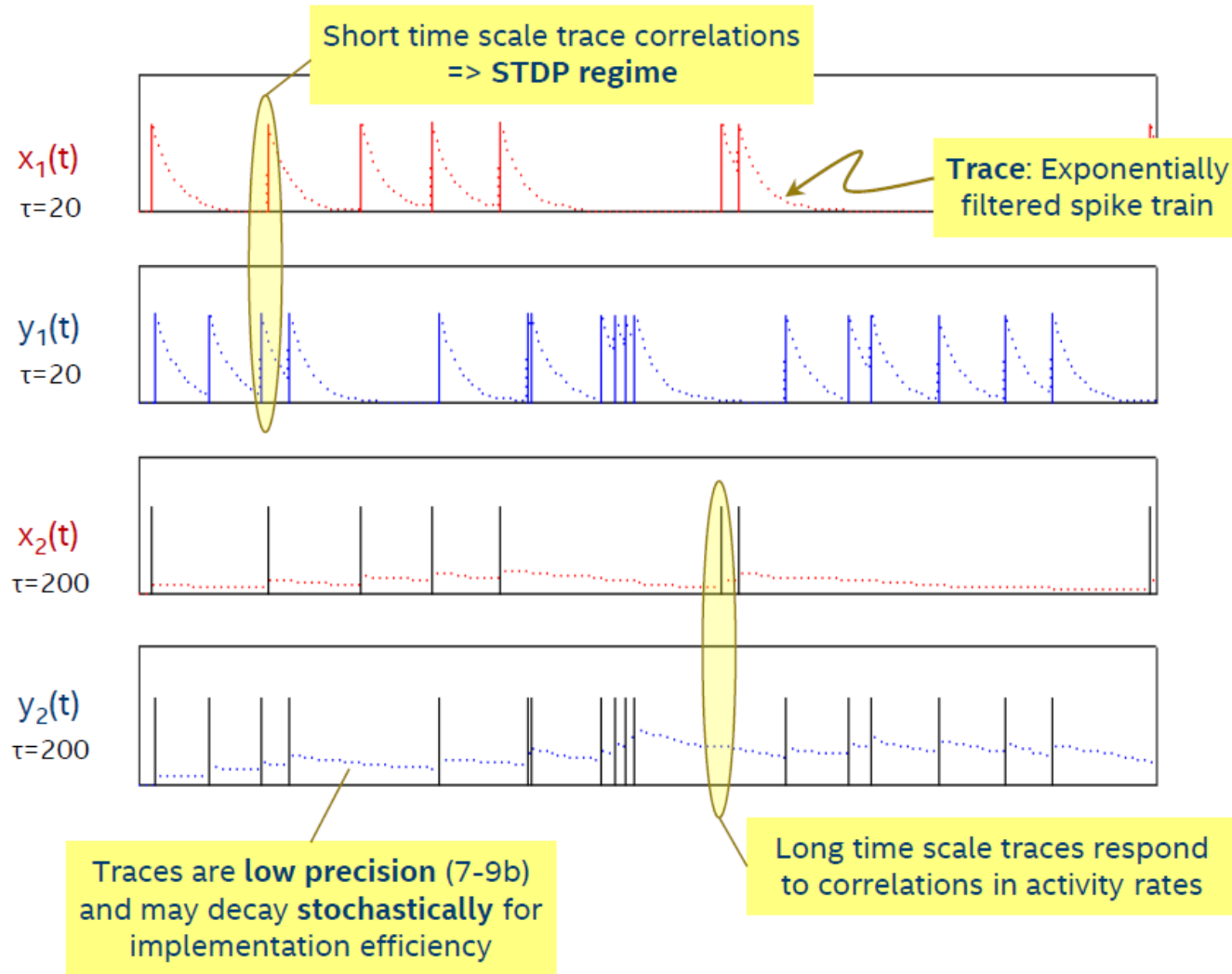
Parallel off-chip interfaces

- Two-phase asynchronous
- Single-ended signaling
- 100-200 MB/s BW

Embedded x86 processors

- Efficient spike-based communication with neuromorphic cores
- Data encoding/decoding
- Network configuration
- Synchronous design

Реализация синаптической пластичности



Weight, Delay, and Tag learning rules programmed as **sum-of-product equations**

$$w' = w + \sum_{i=1}^{N_P} S_i \prod_{j=1}^{n_i} (V_{i,j} + C_{i,j})$$

Synaptic Variables
Wgt, Delay, Tag
(variable precision)

Variable Dependencies
 $X_0, Y_0, X_1, Y_1, X_2, Y_2, R_1$
Wgt, Delay, Tag, etc.

Богатство реализуемых законов пластичности (примеры)

Pairwise STDP:

$$W(t + 1) = W(t) - A_- x_0(t) y_1(t) + A_+ x_1(t) y_0(t)$$

Triplet STDP with heterosynaptic decay:

$$W(t + 1) = W(t) - A_- x_0(t) y_1(t) + A_+ x_1(t) y_0(t) y_2(t) - B \cdot W(t) \cdot y_3(t)$$

Delay STDP:

$$D(t + 1) = D(t) - A_- x_0(t) (127 - y_1(t)) + A_+ (127 - x_1(t)) y_0(t)$$

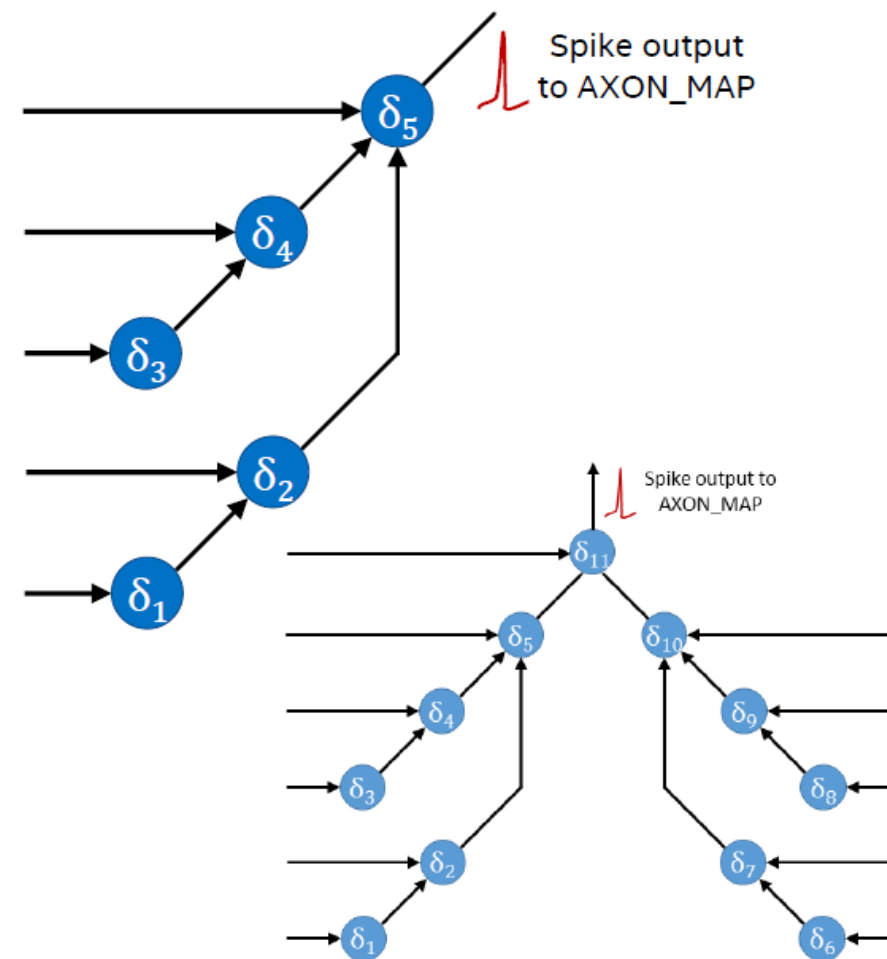
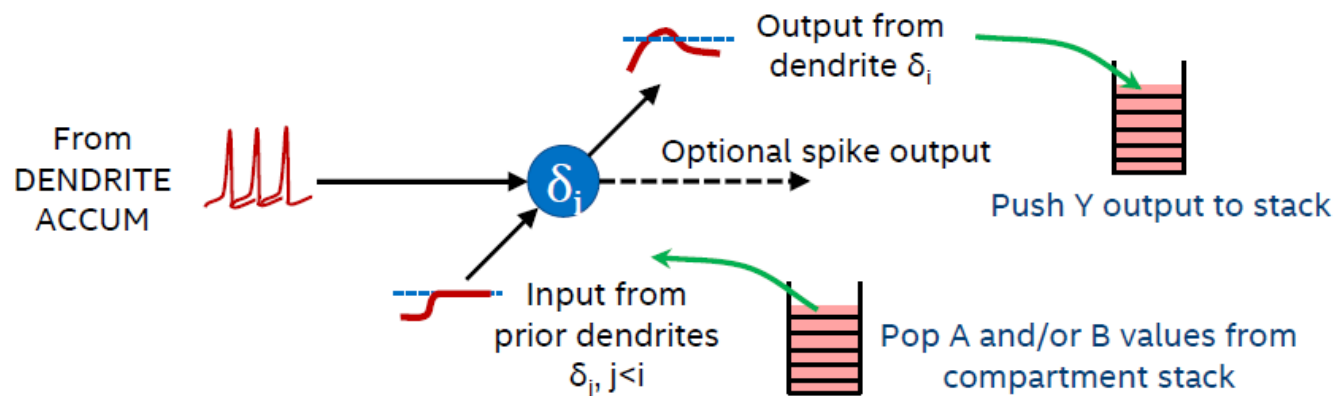
Distal Reward with Synaptic Tags:

$$T(t + 1) = T(t) - A_- x_0(t) y_1(t) + A_+ x_1(t) y_0(t) - B \cdot T(t)$$

$$W(t + 1) = W(t) + C \cdot r_1(t) \cdot T(t)$$

Компонентная модель дендритов – бинарное дерево

Compartment join operations		
0: (NOP)		
1: (ADD_U)	$U' = U + A + B$	
2: (MAX_U)	$U' = \max(U, A, B)$	
3: (MIN_U)	$U' = \min(U, A, B)$	
4: (PASS_U)	$U' = A.S ? U + B : \emptyset$	
5: (BLOCK_U)	$U' = A.S ? \emptyset : U + B$	
6: (OR_S)	$S' = A.S \mid B.S \mid S$	
7: (AND_S)	$S' = A.S \& B.S \& S$	



Нейрокомпьютеры, основанные на Loihi

Q4 2017

Wolf Mountain

Remote Access
4 Loihi/Board



Q2 2018

Nahuku

Arria10 Expansion Board
For cloud & local use
8-32 Loihi/Board



Q3 2018

Kapoho Bay

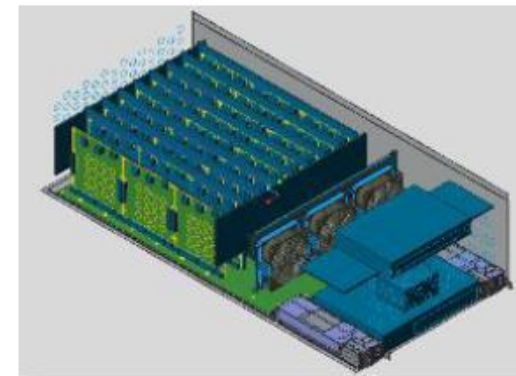
1-2 Loihi
DVS interface
USB host interface



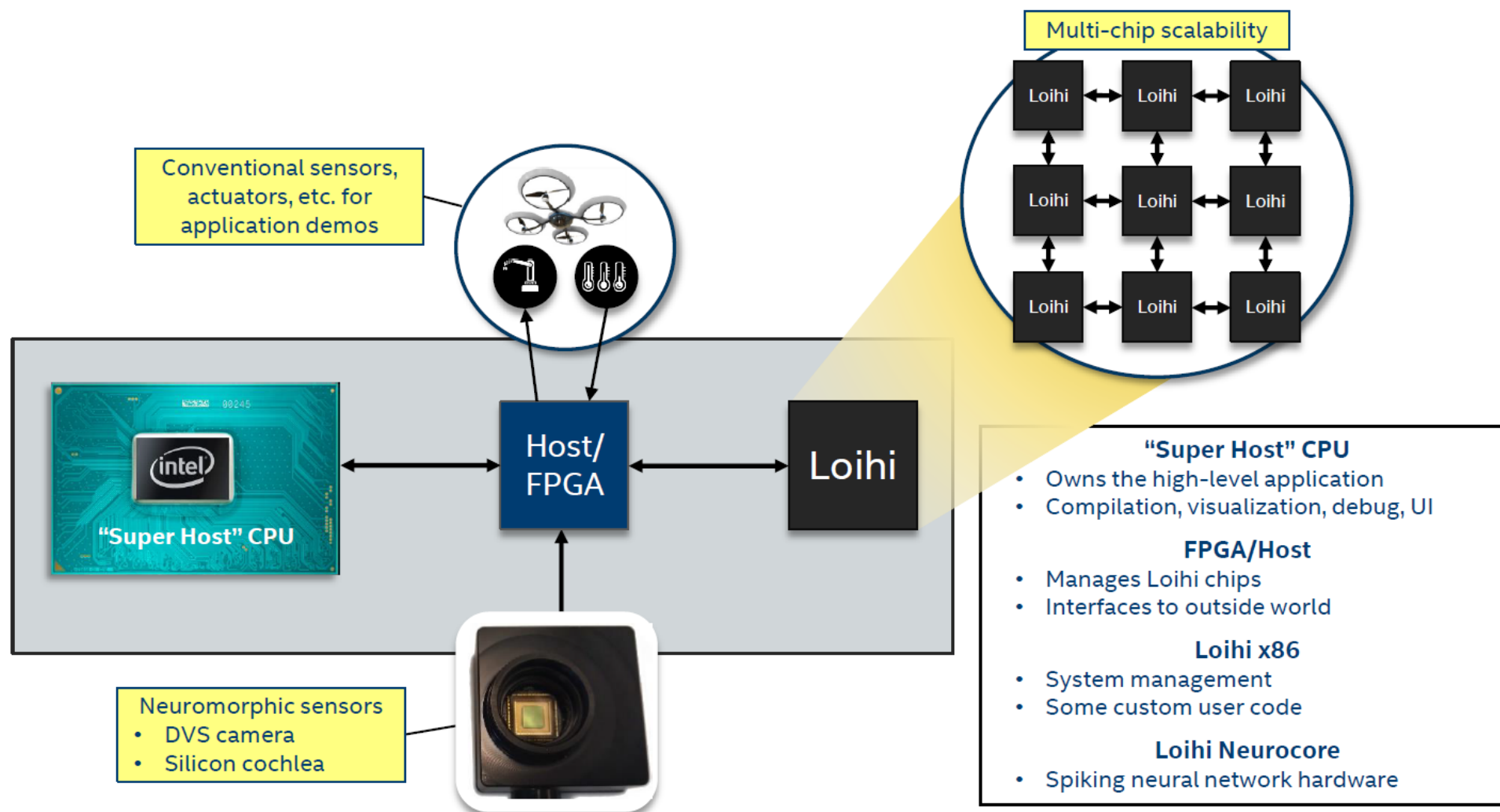
Q2 2019

Pohoiki Springs

Remote Access
Up to 768 chips
(100M neurons)



Общая архитектура системы



Программная архитектура

A module is a complete NxNet defined algorithm (I/O, documentation, etc.)

Computational Modules

LCA	LSNN	EPL	VSA	TPAM
SLIC	CSS	Path Planning	DNF	Astro

3rd party APIs and Frameworks

Nengo	EONS	NRP
ROS	Tensorflow	PyNN

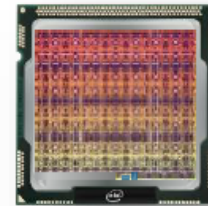
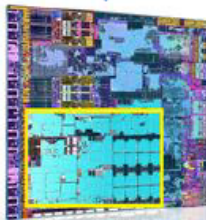
NxNet API

Sequential Neural Interfacing Processes (SNIPs)

Spiking Neural Network (SNN)

NxCompiler

NxCore/NxDriver/NxRuntime



Основные задачи моделирования ИНС в настоящее время

- Развитие методов обучения ИНС (**unsupervised, supervised, reinforcement learning**)
- Моделирование обработки сенсорной информации
- Моделирование механизмов памяти
- Исследование кодирования информации в ИНС и конвертации между разными способами кодирования
- Применение ИНС в системах автоматического управления
- Нейропротезирование и интерфейсы «мозг-компьютер»
- Изучение общих проблем самоорганизации ИНС, теоретических принципов обработки информации ими

Научная тематика лаборатории нейроморфных вычислений ЧГУ (проект ArNI) и лаборатории нейроморфных систем искусственного интеллекта (Цифрум, ЧГУ, Мотив-НТ, «Лаборатория Касперского»)

- разработка моделей импульсных нейронных сетей, имитирующих принципы работы биологических нейронных сетей
- разработка перспективных архитектур нейроморфных систем искусственного интеллекта
- формирование подходов к построению эффективных нейросетевых структур с применением генетических алгоритмов и других эволюционных методов оптимизации
- разработка программных моделей перспективных нейроморфных процессорных элементов, анализ эффективности их реализации в типовых технологических процессах
- изучение различных моделей синаптической пластичности, как базиса для реализации алгоритмов обучения «с учителем и без учителя»
- формирование подходов к созданию нейроморфного процессора с возможностью динамической корректировки синаптических весов (способность к самообучению)

Михаил Киселев, к.т.н.

Проект **ArNI**

Лаборатория нейроморфных вычислений

Чувашский государственный университет

mkiselev@chuvsu.ru

СПАСИБО!

