ヒト大脳皮質の機能を再現する計算機アーキテクチャに関する予備的考察（ミーティング用資料）

2015-09-01

産業技術総合研究所
人工知能研究センター
脳型人工知能研究チーム
一杉裕志、中田秀基

目的

- 脳の機能の再現に特殊なハードウエアが必要だとしたら、早めにハードエウア研究者に予想される要求仕様を伝えておく必要がある。
- そこで、脳を模倣した巨大なニューラルネットワークをシミュレーションするために必要な計算機スペック（演算速度、通信バンド幅、メモリ容量、メモリバンド幅）について、おおまかに考察する。
  - 注：考察の根拠には不確定要素が多い。前提とする数値にもプラスマイナス1ケタ程度の誤差がある。

前提1

- 発火頻度モデルを仮定。
  - もし膜電位レベルのシミュレーションが必要だとすると10年後のスパコンでもおそらく実現不可能。
- 特定の計算モデル（BESOM モデルなど）は陽には仮定しない。極力、神経科学的知見のみを前提として見積もる。

前提2

- ここでは大脳皮質内の回路のシミュレーションだけに注目する。
  - 実際には大脳皮質と大脳基底核、視床、小脳、海馬との相互結合も重要。しかし計算量・通信量は大脳皮質内の回路が支配的であろうと今のところは考えている。
    - 小脳はニューロン数は約 1000億と多いが、その大部分を占める顆粒細胞のシナプス数はあまり多くないらしい（要確認）。

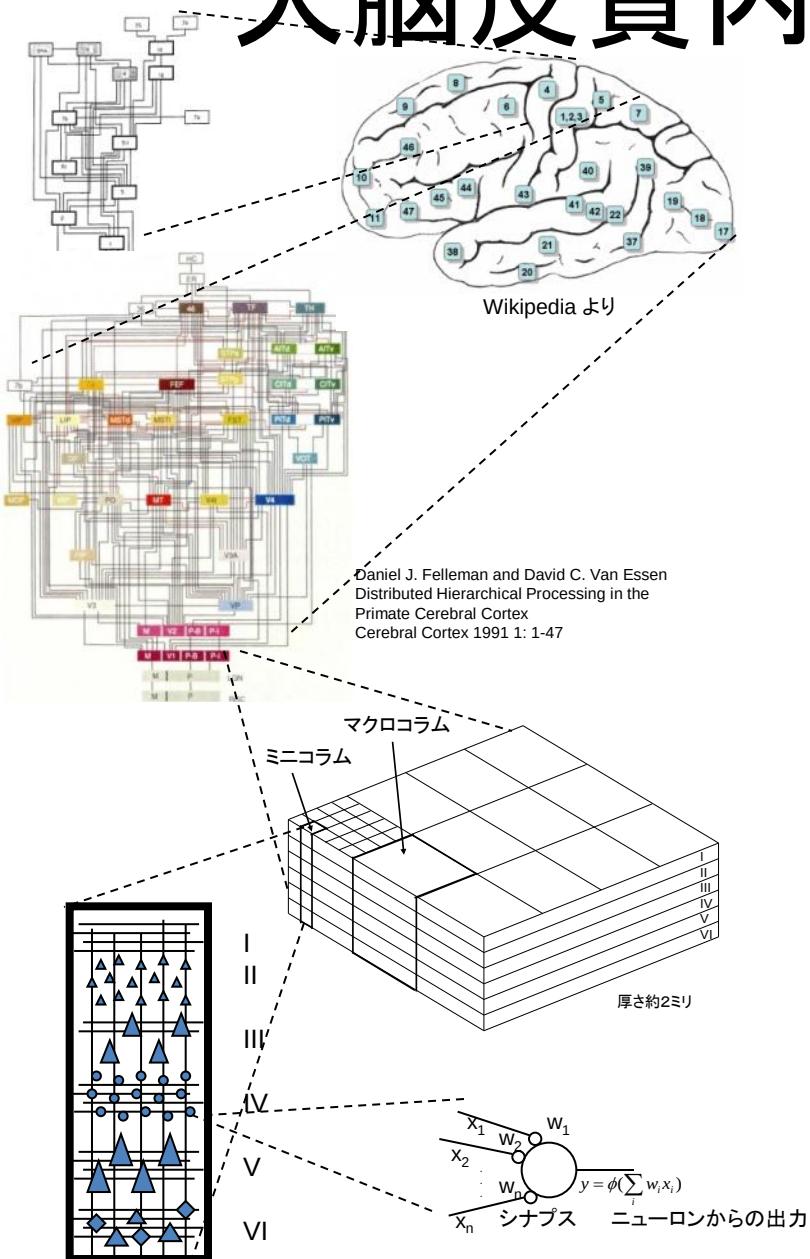
前提3

- シミュレーターの形態：
  1．ハイスペック型
  　脳の動作を実時間より1ケタ以上速い速度で再現、思考・学習を高速化。
  2．実時間型
  　身体を通して外界と実時間でインタラクション。
  3．低コスト普及型
  　機能特化、低速化、低精度化などでコスト削減。
- ここでは2の実時間型を想定。

前提4

- 今回の見積もりは、脳のアルゴリズムを素直に計算機上でシミュレーションする場合を想定。
  - 人工物である計算機上では、生物にはできない最適化アルゴリズムの適用が可能。それにより機能は変えずに何桁も演算数を減らせる可能性がある。
  - その最適化アルゴリズムに適した専用ハードウエアが将来は必要になるかもしれない。
  - しかし、現時点ではどういう最適化アルゴリズムがあり得るのかは不確定。

大脳皮質内の構成要素の個数

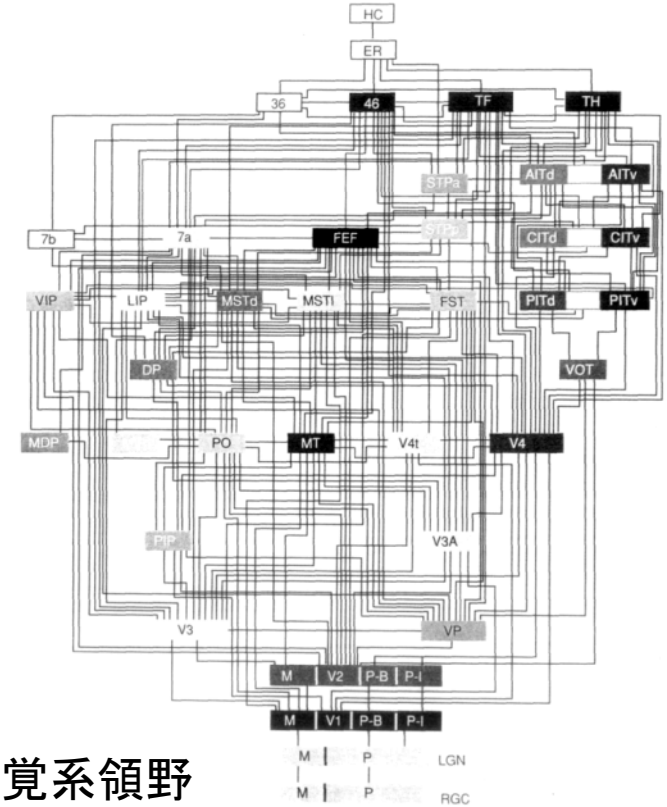


- 領野 100
- マクロコラム100万
  - 1領野あたり1万
- ミニコラム 1億
  - 1マクロコラムあたり100
- ニューロン 100億
  - 1ミニコラムあたり100
- シナプス 100兆
  - 1ニューロンあたり1万

領野間の結合

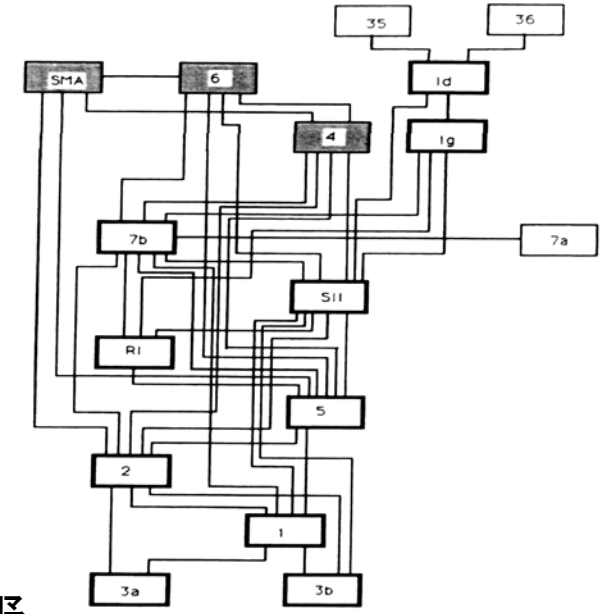
領野のネットワーク

- 完全結合よりはスパースだが、かなり密な結合。
- 領野間の結合は必ず双方向。



視覚系領野

Daniel J. Felleman and David C. Van Essen
Distributed Hierarchical Processing in the Primate Cerebral Cortex
Cerebral Cortex 1991 1: 1-47

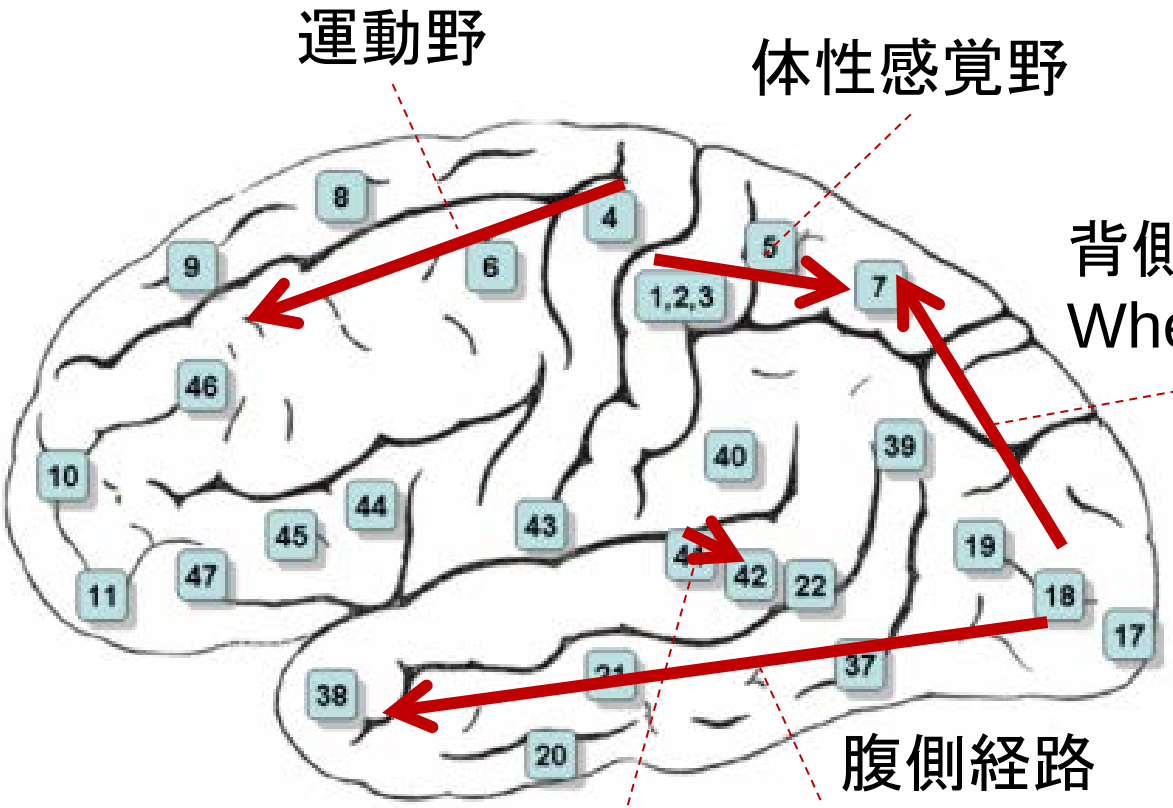


運動野

マクロコラム同士の結合

大脳皮質にみられる領野の階層構造

脳はおそらく様々な感覚入力に対し Deep Learning のような階層的な情報処理を用いている。

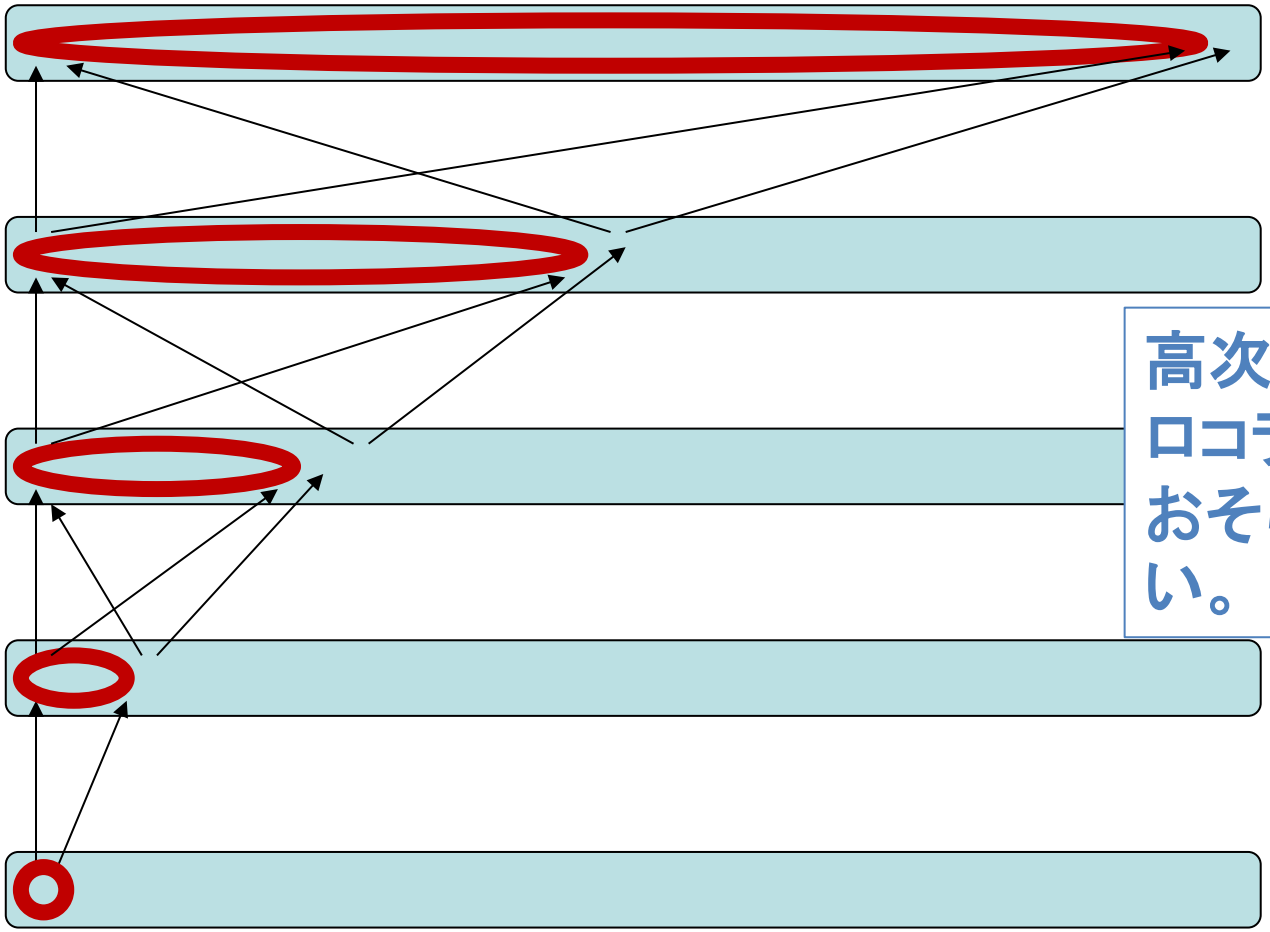


脳の絵は http://en.wikipedia.org/wiki/Brodmann_area より

大脳皮質と deep learning の大きな類似点

- 領野の階層構造
  - 視覚野 腹側経路 V1→V2→V4→PIT→CIT→AIT
  - 視覚野 背側経路 V1→V2→MT・MIP→MST・LIP
  - 体性感覚野　　3野→1野→2野→5野→SII→7b
  - 運動野　　M1→SMA・PM
- Deep learning との大きな類似点
  - **上の層ほど多くの特徴量（感覚入力）を統合。**
  - 上の層ほど複雑な情報を表現。
  - 上の層ほど抽象度が高い情報を表現。

下から上に情報が伝わる範囲



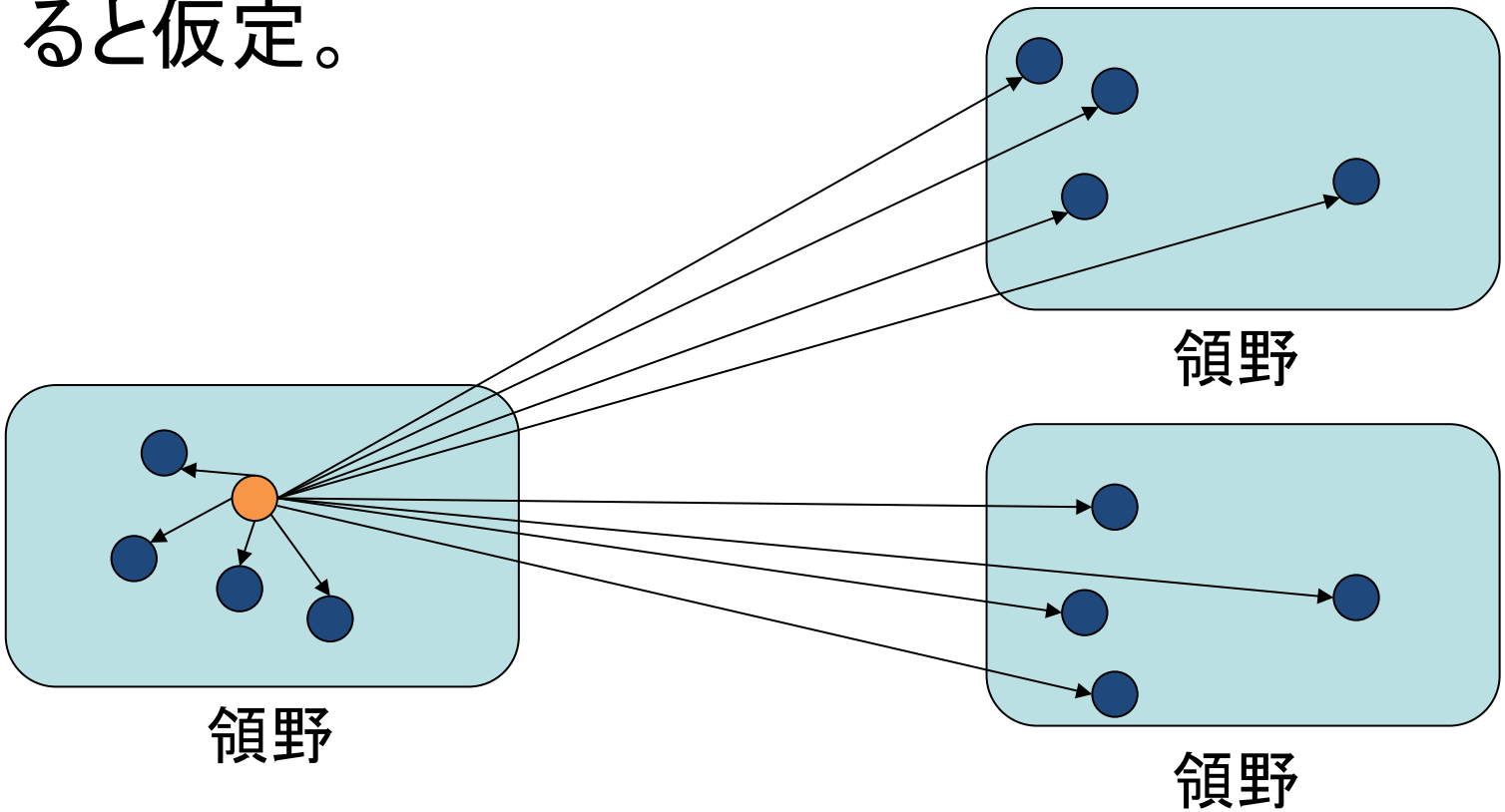
上位層ほど、
広い範囲の中の
どこかに接続

高次領野では、マクロコラム間の結合におそらく局所性はない。

入力層

マクロコラム外への結合

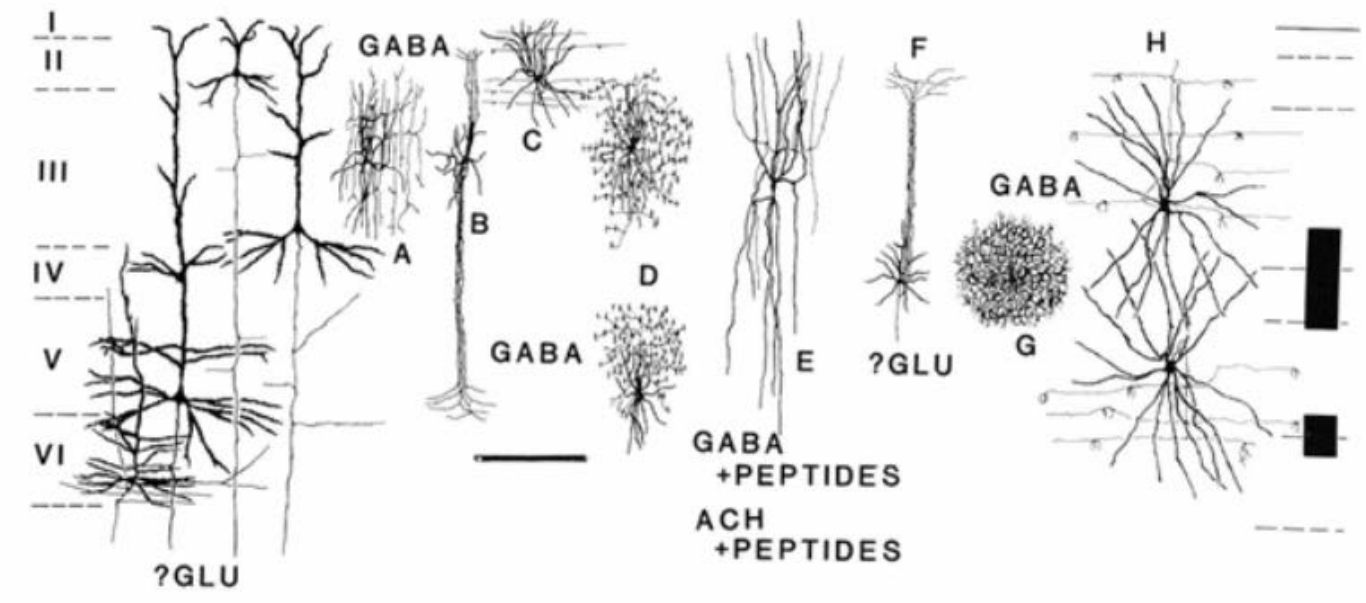
- マクロコラム外に出力するニューロンは、領野の内外の**10個程度**のマクロコラムに結合？
- 明確な根拠は持ってないが、ここでは、そうであると仮定。



マクロコラム内の結合と
マクロコラム外への出力

神経線維のほとんどは マクロコラム（直径500μ）内に局所的

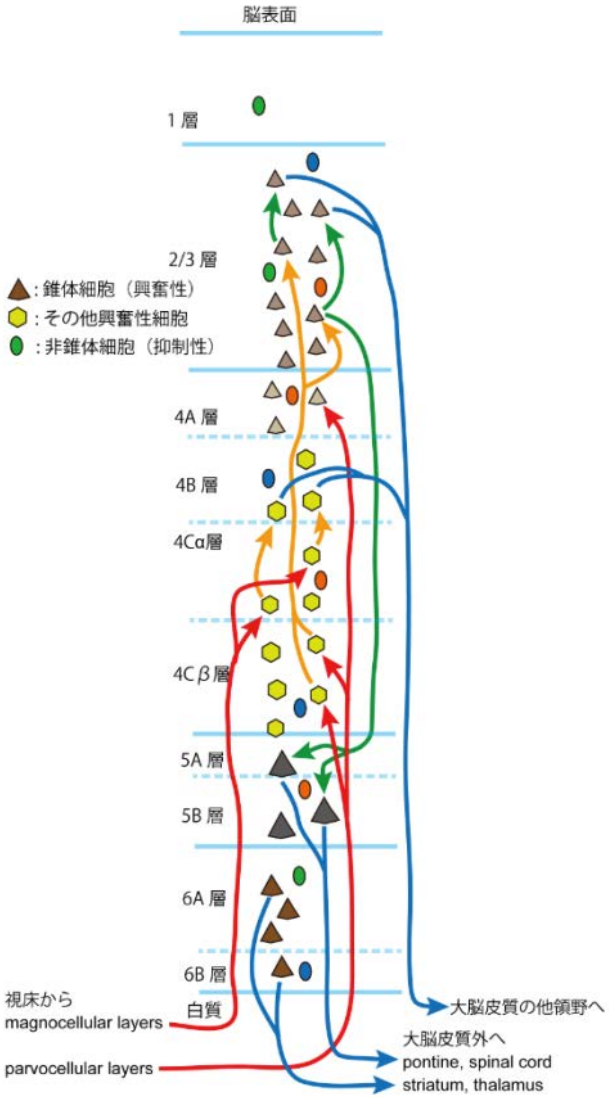
- 樹状突起・軸索の広がりはだいたいマクロコラムと同じオーダー。



Cerebral Cortex
Volume 6 of the series Cerebral Cortex pp 237-266
GABA—Peptide Neurons of the Primate Cerebral Cortex
A Limited Cell Class
E. G. Jones, S. H. C. Hendry, J. DeFelipe

マクロコラムの入出力

- ミニコラム内のニューロン数約100のうち、マクロコラムの外に出力するのは10個程度。
- マクロコラム外から受け取る入力の数もおそらくそのくらい。
- ミニコラムの入出力チャネルをあわせて10個程度と仮定する。
- 1個のマクロコラム内にはミニコラム100個なので、**マクロコラムからの入出力チャネル数は1000個程度。**



ミニコラム内の各ニューロンの役割

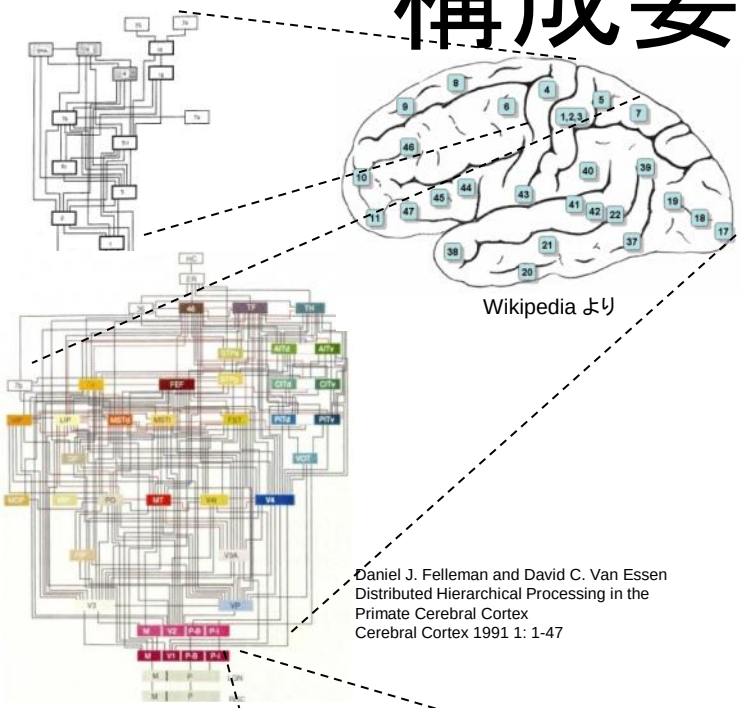
「大脳皮質の局所神経回路 - 脳科学辞典」
https://bsd.neuroinf.jp/wiki/%E5%A4%A7%E8%84%B3%E7%9A%AE%E8%B3%AA%E3%81%AE%E5%B1%80%E6%89%80%E7%A5%9E%E7%B5%8C%E5%9B%9E%E8%B7%AF

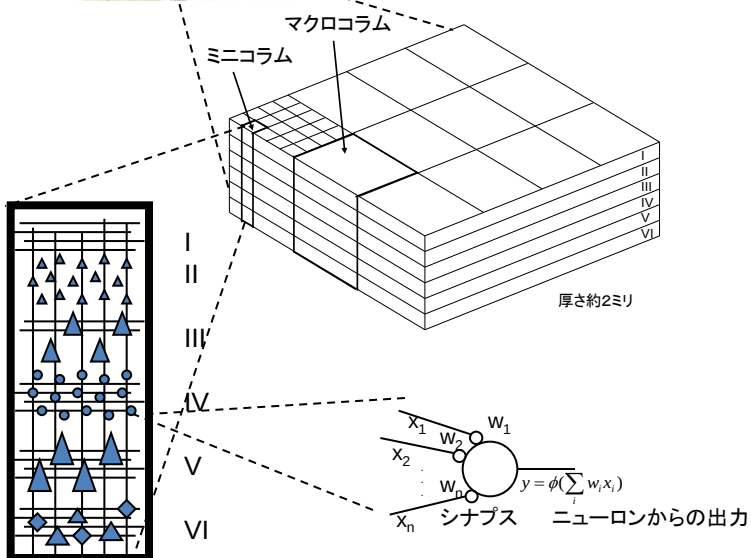
マクロコラム内の結合

- マクロコラム内のニューロン数1万、そのほとんどがマクロコラム内に結合。
  - 1万 x 1万の完全結合に近いと考える。
  - **各ニューロンが同じマクロコラム内のニューロンにブロードキャストするのに近い。**
- 1つのマクロコラムを1つの共有メモリマシンでシミュレーションするのが効率的と思われる。その場合マクロコラム内通信は単なるメモリアクセスとして実現される。

構成要素間の結合のまとめ

計算機のスペックを見積もる際には、領野の接続構造やミニコラム内の構造は重要ではないと考える。





- マクロコラム間結合：
  - ランダムかつ少数
  - 結合先は10個程度
- マクロコラム内のニューロン結合：
  - 完全結合とみなす
- マクロコラムごとの入出力チャネル数
  - 比較的少数
  - 1000個程度

ニューロン間の通信

ニューロン間の通信の特徴

- スパイク最大 1kHz 。発火頻度で情報表現。
- 10ms くらいのオーダーで発火頻度が変化。
- 領野内の演算時間も 10ms 程度。
- 神経線維の信号伝達速度は十分高速。
  - 10m/s = 10cm/10ms以上？
- 発火頻度を 10bit で表現、1秒間に 100 回送るとすると **1K bit/s** のバンド幅。
- 発火はスパース（多くが0に近い値）。

参考: Loopy Belief Propagatoin（確率伝搬法）の特徴

- 有望な仮説:「大脳皮質は loopy BP アルゴリズムを実行している」
  - マクロコラム間での通信を反復するアルゴリズム。
- Loopy BP で行われる通信の特徴:
  - 非同期通信。値がまだ届いていなければ直前に届いた値で代用してもほとんど問題ない。
  - 状態を大きく変える値を優先的に送信すればネットワーク全体の状態が速く収束する。

大脳皮質シミュレーションに必要な計算機アーキテクチャ

大脳皮質の演算速度

- 大脳皮質シミュレーションに必要な演算のほとんどはシナプスが行う入力と重みの掛け算であると仮定。
- 1つのシナプスが 10ms ごとに浮動小数点の掛け算を実行していると仮定すると、脳全体では 100兆 x 100回/秒 = **10P FLOPS**

注：いつも私は1ニューロンあたりのシナプス 1000個で計算して 1P FLOPS と言っていたが、ここでは 10000個という仮定で計算。

大脳皮質のメモリ容量(1/4)

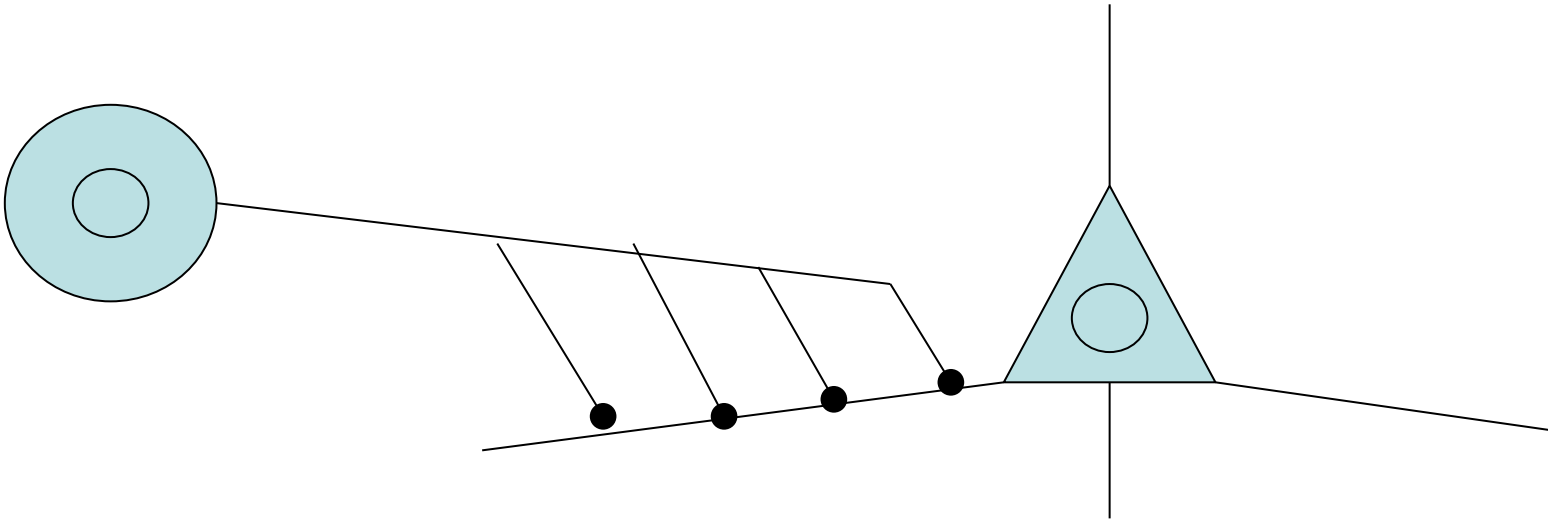
- 長期記憶:シナプス100兆 x シナプス1個あたり 1byte とすると 100TB
  - しかしこれは過大すぎる見積もり。後述。
- マクロコラム間の通信バッファ
  - マクロコラムの数100万 x 通信チャネル1000個 x 1byte とすると 1GB
- 1つのマクロコラム内の通信バッファ
  - マクロコラム内のニューロン数1万 x 1byte とすると 10KB

大脳皮質のメモリ容量(2/4)

- シナプスのすべてが可塑性シナプスというわけではない。
- マクロコラム間の結合のみが可塑性シナプスであると仮定する。
- マクロコラム内の1万個のニューロンのうち1000個が入出力チャネルという仮定なので、可塑性シナプスも 1/10 と仮定する。

大脳皮質のメモリ容量(3/4)

- あるニューロンの出力は同じニューロンに重複して結合する。結合するシナプスの総和で1つの重みを冗長に表現。
- 冗長度 1/100 と仮定する。



大脳皮質のメモリ容量(4/4)

- 長期記憶:シナプス100兆 x 可塑性シナプスの比率 1/10 x 冗長度 1/100 x シナプス1個あたり 1byte とすると **100GB**
  - 注: 8bit のシナプス100個の総和で表現できる値は 0-25500 で 15bit 程度の情報量だが、概算なので 1byte とした。

必要なメモリバンド幅

- シナプスがメモリ。演算速度と同程度の速度でアクセスできる必要がある。
  - ほとんどメモリアクセスがレジスタか L1 cache 程度の速度でアクセスできないといけない。
- 全体で必要なメモリバンド幅は、
  メモリ容量 100GB x 1秒当たり100回の演算
  = **10TB/s**

ランニングコスト

- 電気代 20円 / 1kWh 、人間の雇用コスト 2000円/h とすると、人間の雇用コストと等価な電力は 100kW
- 大脳皮質をそのコストで置き換えるために必要な電力効率は、10P FLOPS / 100kW = 10,000,000 G FLOPS / 100 kW
  = **100G FLOPS / W**
- 今年の Green500 トップの PEZY のマシンは 7G FLOPS / W

見積もり1

- １計算機ノード **100G FLOPS** とする。
  - 最近の Xeon はそのくらい。
- 10 P FLOPS では 100,000 計算機ノードが必要。
  １計算機ノードあたり10マクロコラムを担当。
- 通信バンド幅は１計算機ノードあたり
  10マクロコラム x 10個の接続先 x 1000チャネル x 1K bit/s
  = **100M bit/s**
- メモリ １計算機ノードあたり 100GB/100,000 = **1MB**
- メモリバンド幅 １計算機ノードあたり
  10TB/100,000 = **100MB/s**
- １計算機ノード１００万円として**１０００億円**。

見積もり2

- １計算機ノード **1T FLOPS** とする。
  - GPGPU を想定
- 10 P FLOPS では 10,000 計算機ノードが必要。
  １計算機ノードあたり１００マクロコラムを担当。
- 通信バンド幅は１計算機ノードあたり
  100マクロコラム x 10個の接続先 x 1000チャネル x 1K bit/s = **1G bit/s**
- メモリ １計算機ノードあたり
  100GB/10,000 = **10MB**
- メモリバンド幅 １計算機ノードあたり 10TB/10,000 = **1GB/s**
- １計算機ノード１００万円として**１００億円**。
  - ノード間の通信機器の価格もおそらく同じくらいのオーダー。

考察1

- GPGPU クラスタを用いると割とちょうどよいバランス。
- メモリバンド幅は DRAM で十分。キャッシュが理想的に効くなら問題ない。効かないならそこがボトルネックになり得る。
- 現状の GPGPU はコア間の通信を高速に行う機構がなく、そこも問題になる可能性がある。

考察2

- **労働の自動化という社会要請に応えるために必要なのは低コスト化。**
- ２０年後には半導体集積度向上によりチップのコストは 1/100 に？
- 計算機ノードをマクロコラムの実行に最適化すればさらに 1/10 にすることが可能か？
- そうであれば、２０年後にコスト 1/1000 に。１００億円が１０００万円に下がる。労働力として競争力がある値段に。

最後に

- 今回の見積もりは、まだかなりあやしい。
- いまのところは、脳のシミュレーションにそれほど特殊なバランスのハードウエアは必要なさそう。
- 間違いなく必要なのは、初期コストとランニングコストの両方の大幅な低コスト化。

以上

参考:
近似確率伝搬アルゴリズムと
大脳皮質のコラム構造との対応

近似確率伝播アルゴリズム[Ichisugi 2007]

$$\boldsymbol{l}_{XY}^{t+1} = \boldsymbol{z}_Y^t + \boldsymbol{W}_{XY}\boldsymbol{o}_Y^t$$

$$\boldsymbol{o}_X^{t+1} = \prod_{Y \in children(X)}^{\otimes} \boldsymbol{l}_{XY}^{t+1}$$

$$\boldsymbol{k}_{UX}^{t+1} = \boldsymbol{W}_{UX}^T \boldsymbol{b}_U^t$$

$$\boldsymbol{p}_X^{t+1} = \sum_{U \in parents(X)} \boldsymbol{k}_{UX}^{t+1}$$

$$\boldsymbol{r}_X^{t+1} = \boldsymbol{o}_X^{t+1} \otimes \boldsymbol{p}_X^{t+1}$$

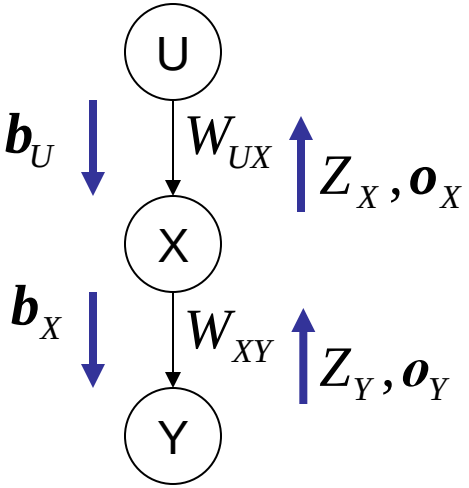
$$Z_X^{t+1} = \sum_i (\boldsymbol{r}_X^{t+1})_i \ (= \left\| \boldsymbol{r}_X^{t+1} \right\|_1 = \boldsymbol{o}_X^{t+1} \bullet \boldsymbol{p}_X^{t+1})$$

$$\boldsymbol{z}_X^{t+1} = (Z_X^{t+1}, Z_X^{t+1}, \cdots, Z_X^{t+1})^T$$

$$\boldsymbol{b}_X^{t+1} = (1/Z_X^{t+1})\boldsymbol{r}_X^{t+1}$$

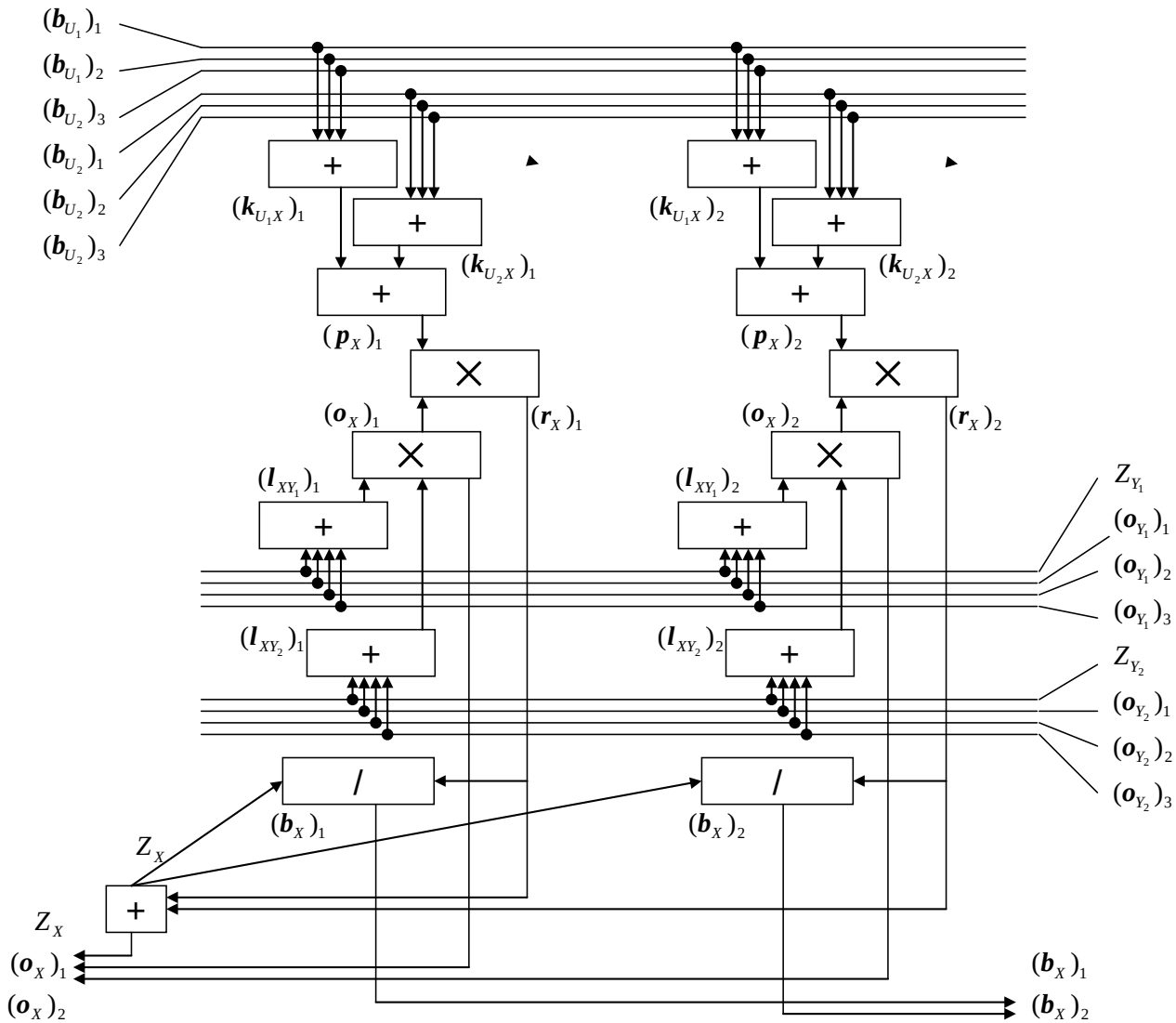
ただし、$\boldsymbol{x} \otimes \boldsymbol{y} = (x_1 y_1, x_2 y_2, \cdots, x_n y_n)^T$

Pearl のアルゴリズム[Pearl 1988]をいくつかの仮定のもとで近似。
（アルゴリズムは今後少し修正する予定）



・神経回路で実現可能
・大規模化可能な計算量・記憶量

各変数の値を計算する回路



I

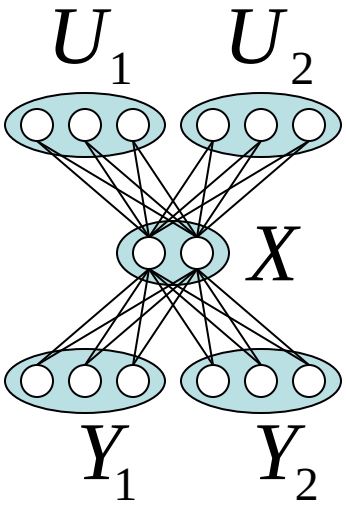
II

III

IV

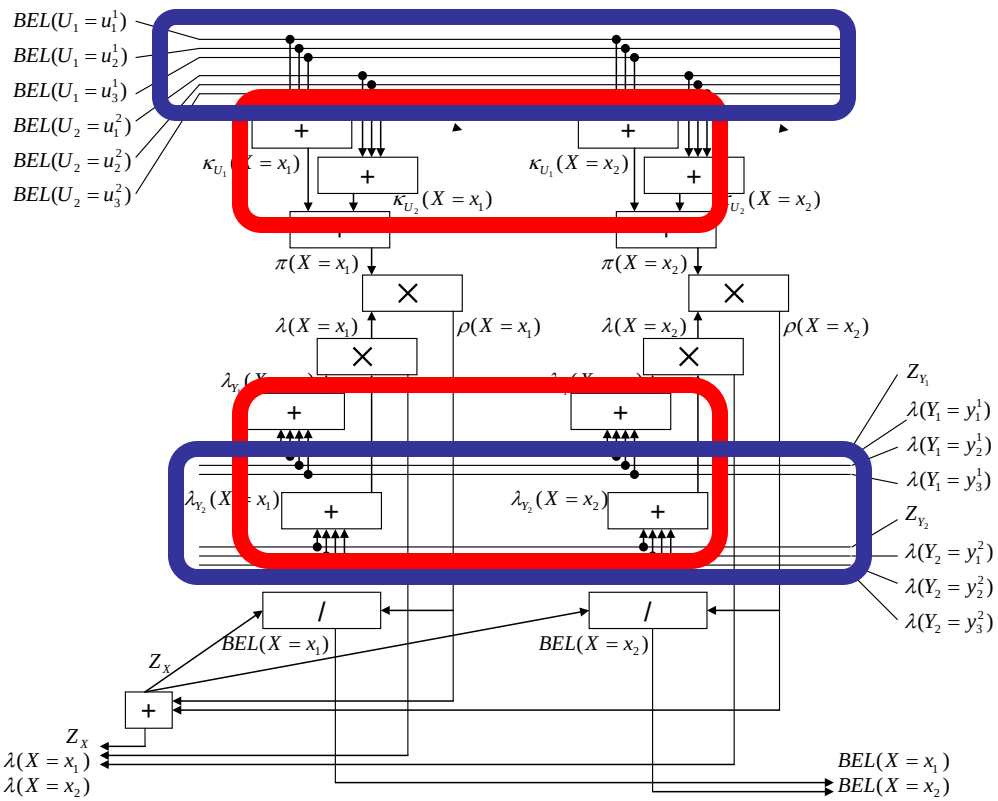
V

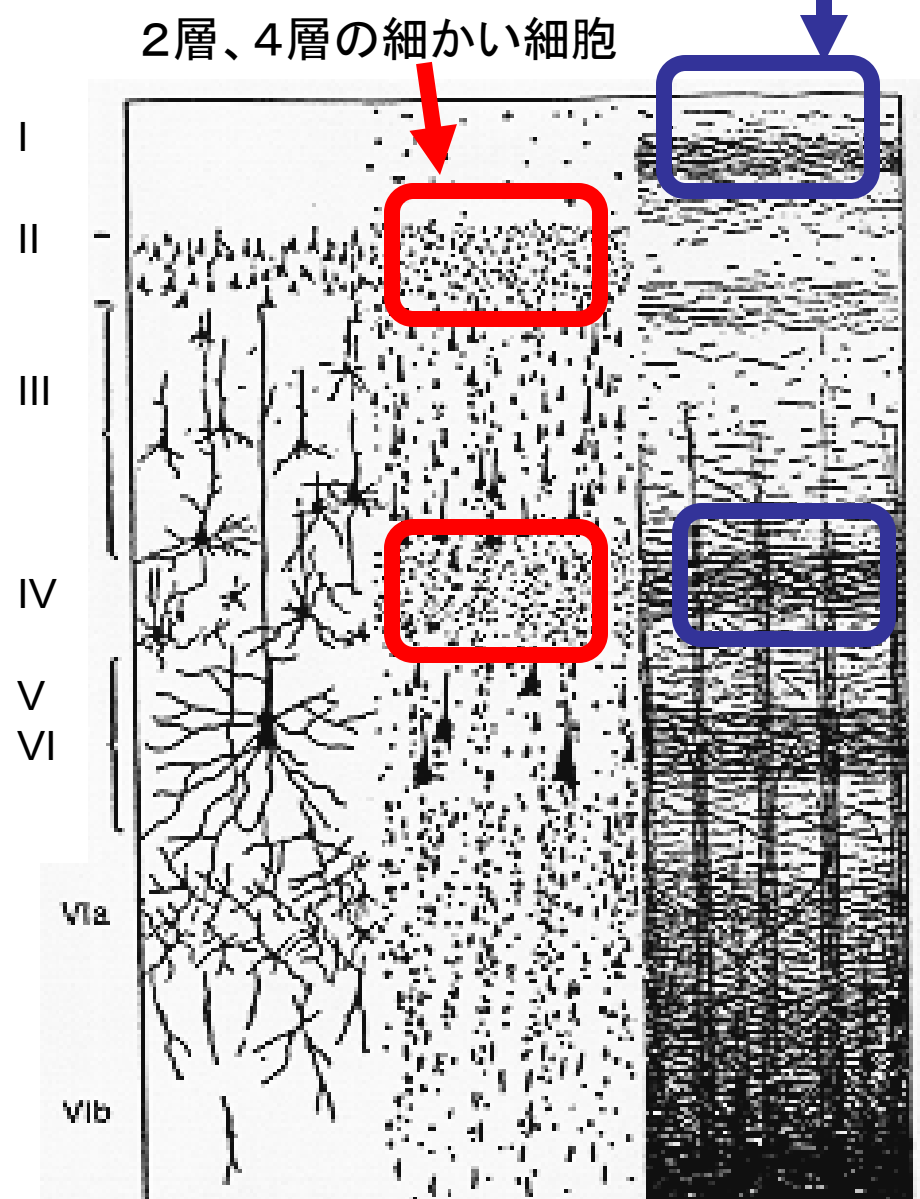
VI



左図は、上記BESOMネットワークのノードXのユニットの値を計算する回路

大脳皮質の構造との一致





K. Brodmann, Vergleichende Lokalisation der Grosshirnrinde. in: ihren Prinzipien dargestellt auf Grund des Zellenbaues,. J.A. Barth, Leipzig, 1909.

This figure is taken from the following Web page.
http://web.sc.itc.keio.ac.jp/anatomy/brodal/chapter12.html

想定される BESOM ネットワークの特徴

- １階層を上がるごとに受容野が広がる
（だいたい面積が４倍くらい）
- 親子ノード間の結合は必ず双方向。
- あるノードから見た親ノードの数、子ノードの数は一定
（４～１０個くらい）
- どの層もノード数はだいたい一定。
（ヒトで１領野１万ノードくらい）
- 上の層ほど発火はスパース。
- 層内の結合（ノード間の側抑制）もある。おそらく1ユニットあたり１０個程度ですむ。