KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
PROGRAMŲ INŽINERIJOS KATEDRA

Justas Sasnauskas

Neuroninio procesoriaus prototipas FPGA technologijoje

Magistro darbas

Vadovas
prof. habil. dr. E. Kazanavičius

KAUNAS, 2005

SUMMARY

In this paper it is described a method of creation of a neuroprocessor prototype, from high abstraction level to FPGA technology. Most common neuroprocessor architectures are overviewed, and canonical model of the neuroprocessor is created accordingly. Based on the canonical model the serial structure neuroprocessor mathematical model is formed and evaluated. The model is than described in SystemC programming language. In the experimental part, the correct functionality of the neuroprocessor is evaluated and the results of synthesis are analyzed.

Turinys

1. Įvadas

Augant informacijos srautams ir plečiantis informacinių sistemų taikymų įvairovei atsiranda vis didesnis poreikis intelektualiam duomenų apdorojimui. Klasikiniais skaičiavimo metodais sunku vertinti nuolat kintančius aplinkos parametrus, juos interpretuoti, apdoroti. Paskutiniais dešimtmečiais atsiradus DNT (dirbtiniams neuroniniams tinklams) tokie uždaviniai tapo išsprendžiami. Biologinių neuroninių tinklų principais sukurti DNT pasižymi lankstumu, aukštu autonomijos laipsniu, gebėjimu interpretuoti ir vertinti duomenis nepriklausomai nuo jų tikslumo. Viena iš svarbiausių neuroninių tinklų savybių yra gebėjimas mokytis. Tai leidžia kurti sistemas gebančias prisitaikyti prie kintančių sistemos parametrų. Kombinuojant neuroninius tinklus, miglotąją (angl. *fuzzy*) logiką, genetinius algoritmus ir klasikinius skaičiavimo metodus galima pasiekti ypač gerų rezultatų [1].

Neuroniniai tinklai programinėje įrangoje taikomi gana plačiai optinio raidžių atpažinimo, balso atpažinimo ir atkūrimo iš teksto uždaviniams spręsti, autopilotams, žaidimų programoms, grafiniam vaizdui apdoroti ir kt. Tačiau greitiems didelių DNT skaičiavimams atlikti bendros paskirties procesorių neužtenka. Imlūs techniniams resursams DNT skaičiavimai aparatūrinėje įrangoje sunkai įgyvendinami, o esantys taikymai nepasižymi didele komercine sėkme. Dėl plačios neuroprocesorių architektūrinės įvairovės projektuotojai DNT negali universaliai taikyti [2].

Šio darbo tikslas yra pasiūlyti apibendrintą neuroprocesoriaus modelį ir patikrinti pasiūlyto modelio veikimą, realizuojant nuoseklios architektūros neuroprocesoriaus prototipą FPGA (angl. *Field-Programmable Gate Array* – programuojamųjų sklendžių matrica) technologijoje.

Šiame darbe yra apžvelgiamas neuroprocesorių architektūros literatūroje, suformuojamas apibendrintas neuroprocesoriaus modelis, jo pagrindu suformuojamas nuoseklios architektūros neuroprocesorius, kuri užrašomas SystemC programavimo kalba. Eksperimento metu neuroprocesoriaus prototipas realizuojamas Xilinx XC2V40 FPGA įrenginyje, analizuojami sintezės rezultatai.

2. Analitinė dalis

Žmogaus smegenų tyrimai trunka jau tūkstančius metų, tad nenuostabu, kad naudojant atsiradusią modernią elektroniką stengiamasi imituoti paslaptingus mąstymo procesus. Pirmieji sukūrę elektrinę grandinę, imituojančią smegenų veiklą, 1943 m. buvo neurofiziologas (angl. *neurophysiologist*) Warren McCulloch ir matematikas Walteras Pittsas. Kompiuteriams tapus pakankamai galingiems tapo įmanoma modeliuoti įvairias teorijas aprašančias smegenų veikimo principus. Nathanial Rochester IBM tyrimų laboratorijoje 1950 m. buvo atliekami pirmieji modeliavimo bandymai.

Neurobiologas F. Rosenblattas 1958 m., tyrinėdamas vabzdžių akis, sukūrė viensluoksnį neuroninį tinklo modelį, kurį pavadino perceptronu. Šis tinklas buvo išbandytas technikoje ir yra pirmasis dabar naudojamas neuroninis tinklas. Viensluoksnis perceptronas pasižymi informacijos klasifikavimo savybėmis ir gali išmokti bet kokią funkciją kurios išėjimų aibę sudaro du skaitmenys [3][4].

Bernardas Widrowas ir Marcianas Hoffas 1959 m. sukūrė neuroninius modelius kuriuos pavadino ADALINE ir MADALINE (angl. *Multiple ADAptive LINear Elements*). Šie tinklai dar vadinami adaptuojamaisiais filtrais (angl. *Adaptive filter*). Šis modelis buvo pritaikytas aidui telefono linijose panaikinti. Tai buvo pirmasis neuroninis tinklo praktinis panaudojimas konkrečiam uždaviniui spręsti. Tinklų ADALINE išėjimas gali turėti bet kokią reikšmę, tačiau kaip ir perceptronas gali spręsti tik tiesiškai atskiriamas problemas.

Plačiau neuroniniai tinklai pradėti taikyti tik paskutiniame dešimtmetyje atsiradus pajėgesniems asmeniniams kompiuteriams. Didėjant bendros paskirties procesorių pajėgumams dauguma projektuotojų ne sudėtingus, greičio nereikalaujančius uždavinius realizuoja programinėje įrangoje. Tačiau net ir gerokai galingesni nuoseklūs bendros paskirties procesoriai negali skaičiuoti ir apmokyti didelių neuroninių tinklų su dideliu neuronų ir sinapsių skaičiumi. Čia išeitis gali būti neuroprocesoriai, turintys lygiagrečiai veikiančius paprastus komponentus.

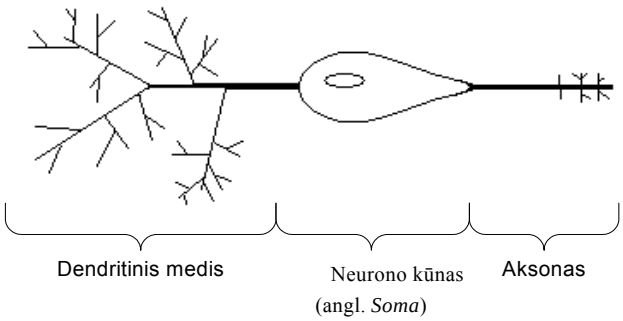
Ligi šiol DNT aparatūrinės įrangos projektuotojai rinkosi vieną iš dviejų skirtingų projektavimo krypčių: kurti sudėtingus brangius plačios paskirties neuroprocesorius arba paprastus ir greitus specifiniam uždaviniam skirtus lustus. Skirtingi NP projektuotojai randa skirtingus atsakymus į šiuos klausimus [5-6].

2.1. Neuroniniai tinklai

Dirbtinis neuroninis tinklas tai supaprastintas matematinis biologinio neuroninio tinklo modelis imituojantis tarpusavyje susijungusias ląsteles, vadinamas neuronais, ir atliekantis tam tikrus smegenyse vykstančius procesus.

Kiekvienas neuronas su kitais neuronais sudaro tūkstančius sinapsių, o kadangi žmogaus CNS (centrinėje nervų sistemoje) yra apie 10^{11} neuronų, tai bendras sinapsių skaičius siekia 10^{14}, taigi žmogaus smegenyse sinapsių yra daugiau negu sakykime, žvaigždžių mūsų galaktikoje [6].

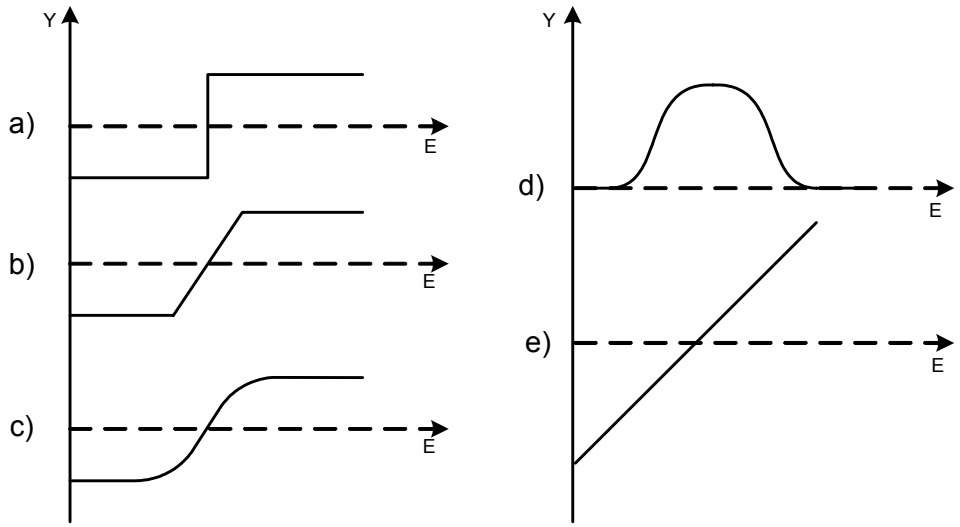
Biologinių neuronų būna įvairių formų ir paskirčių, tačiau visi jie turi tris pagrindines dalis dendritinę, aksoninę ir somatinę. Dendritais neuronas prisijungia prie netoliese esančių kitų neuronų, sudarydami jungtis, vadinamas sinapsėmis. Tokių jungčių kiekis neurone yra nepastovus ir kinta atsižvelgiant į neurono padėtį ir paskirtį organizme. Pavyzdžiui, CNS (centrinėje nervų sistemoje) neuronas turi apie du tūkstančių jungčių. Dėl dauginės neurono ląstelės sužadinimo kiekvienas aksonas sukelia palyginti mažą posinapsinį potencialą, tuo tarpu visi kartu, visada viršija sužadinimo slenkstį [6]. Sužadintas neuronas toliau siunčia signalą aksonu kitiems prie jo prisijungusiems neuronams. Nesužadintas neuronas signalo nesiunčia, tačiau sužadintas kurį laiką siunčia didelio potencialo signalą.



1 pav. Biologinis neuronas

Šiomis savybėmis remiasi ir dirbtiniai neuronų modeliai. Neuronai nereaguoja į per mažą suminį potencialą, o įsisotinęs nereaguoja į per didelį. Tai yra neuronas jautriausias tada, kai įėjimų suminis signalas svyruoja ties sužadinimo slenksčiu.

Remiantis biologinių neuronų veikimo principais pirmieji McCulloch-Pitto (1943m.) sudaryti matematiniai neurono modeliai turėjo slenkstinę perdavimo funkciją [2]. Tačiau norint pasiekti neriboto pločio signalą praktikoje naudojamos ir tiesinės funkcijos, kurios paprastai naudojamos neuroninio tinklo išėjimo sluoksniuose. Dažniausiai pasitaikančios dirbtinių neuronų perdavimo funkcijos pavaizduotos 2 paveiksle.

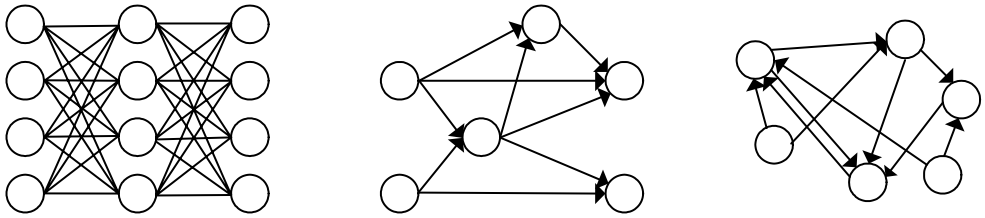


2 pav. Dažniausiai sutinkamos dirbtinių neuroninių tinklų perdavimo funkcijos a) slenkstinė, b) tiesinė ribojanti, c) hiperbolinio tangento d) radialinės bazės e) tiesinė

Per visą neuroninių tinklų plėtojimo laikotarpį buvo sukurtas ne vienas neuroninių tinklų modelis, pasižymintis skirtingomis veikimo savybėmis. DNT skirstomi pagal:

- signalo sklidimo kryptį:
  - tiesioginio sklidimo neuroninis tinklas
  - grįžtamojo ryšio neuroninis tinklas
- neurono perdavimo funkciją;
- neuroninio tinklo paskirtį:
  - filtravimas, aproksimavimas
  - informacijos klasifikavimas
  - analizė
- mokymo būdą:
  - mokymas su mokytoju
  - mokymas be mokytojo

Pagal signalo sklidimo kryptį neuroniniai tinklai skirstomi į dvi grupes: tiesioginio sklidimo (TS) (angl. *feed-forward*) ir grįžtamojo ryšio (angl. *recurrent*) neuroninius tinklus (3 pav.). Tiesioginio sklidimo neuroninių tinklų neuronai sujungti taip, kad neuronais sklindantis signalas niekada nesudarytų uždarų grandinių. Grįžtamojo ryšio neuroniniai tinklai (GRNT) gali būti sujungti bet kokia tvarka. Praktikoje GRNT plačiai netaikomi dėl savo sudėtingumo. Be to, GRNT dar nėra gerai ištirti, nėra sukurtų patikimų mokymo algoritmų [2].



3 pav. a)tiesioginio sklidimo sluoksninis tinklas (angl. *Layered feed forward neural network*); b)tiesioginio sklidimo neuroninis tinklas (angl. *Feed forward neural network*) c)grįžtamojo ryšio neuroninis tinklas (angl. *Recurrent neural network*)

Tiesioginio sklidimo neuroniniai tinklai pasižymi:

- funkcija, statiškai siejančia įėjimus su išėjimais;
- daugeliu praktinių taikymų, susijusių su funkcijų aproksimacija ir informacijos klasifikavimu;
- galimomis mažesnėmis nei 10 bitų pločio duomenų struktūromis;

Grįžtamojo ryšio neuroninis tinklas pasižymi:

- panašumu su biologiniais tinklais, nes visi biologiniai tinklai yra grįžtamojo ryšio;
- galimybe matematiškai įgyvendinti dinamines sistemas;
- sudėtingumu ir brangumo;

Uždaviniams, kai reikia nuolatinės laike užvėlintos duomenų sekos analizės, sluoksniniai neuroniniai tinklai yra netinkami, o paprasti tiesioginio sklidimo neuroniniai tinklai gali tokius uždavinius spręsti tik iš dalies, sudarydami neuronines informacijos sekas duomenų kitimui analizuoti. GRNT pasižymi ne tik ilgalaike atmintimi (angl. *Long-term memory*), formuojama keičiant svorius, bet ir trumpalaike (angl. *Short-Term Memory*), kuri susidaro dėl grįžtamųjų ryšių [7]. Dėl šios savybės GRNT

gali ne tik analizuoti skirtingo laiko įvykius, bet ir būti generatoriais, formuojančiais įvairias tolygiai ar drastiškai kintančias dinamikas [8].

Neuroprocesoriai, galintys apdoroti GRNT, gali apdoroti ir TSNT, dėl to jie gali būti pritaikomi plataus spektro uždaviniams spręsti.

2.2. Neuroprocesorinių struktūrų apžvalga

Aparatūrinių sprendimų neuroninių tinklų skaičiavimui įvairovė gana plati. Nuo paprastų mikrokontrolerių iki galingų neurokompiuterių. Neuroniniai tinklai realizuojami analoginėje, skaitmeninėje, hibridinėje elektronikoje, moksliniai tyrimai atkiekami optoelektronikos (angl. *optoelectronics*), molekulinės elektronikos (angl. *molecular* electronics), hibridinių biolustų (angl. *hybrid biochips*) srityse. Neuroninių sistemų daug, o pasirinkti sunku. Remiantis [2] [9], neuroprocesorius galima skirstyti pagal:

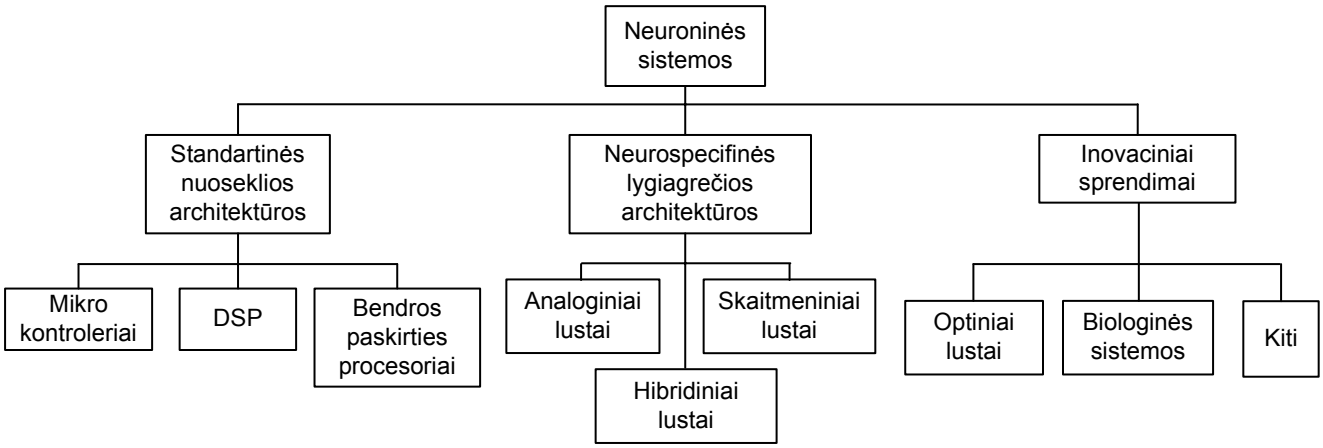
- sistemos architektūrą
- lygiagretumo laipsnį
- pasikeitimo duomenimis architektūrą
- paskirtį (specifinės ar bendros)
- mokymo tipą (vidinis ar išorinis)
- svorių, įėjimų bei išėjimo duomenų plotį
- neurono perdavimo funkciją
- kainą
- energijos suvartojimą
- kt.

Pagal sistemos sandarą neuroninė aparatūrinė įranga skirstoma į tris pagrindines architektūras: standartines nuosekliąsias architektūras, neurospecifines lygiagrečias architektūras ir inovacines architektūras (4 pav.).

Standartiniams sprendimams priklauso bendrosios paskirties procesoriai, pritaikyti neuroniniams tinklams skaičiuoti. Ši klasė taikoma neurokompiuteriuose, neuroninių tinklų skaičiavimus spartinančiose plokštėse, įterptinėse sistemose. Neuroninių tinklų aparatūrinės įrangos projektuotojai renkasi juos dėl palyginti mažos kainos. Jie tinka tokiems uždaviniams spręsti, kuriems nereikia apdoroti didelio informacijos srauto. Procesorius jungiant lygiagrečiai, ypač tuos, kurie atlieka labai greitus sandaugos, sumos kaupimo operacijas, galima pasiekti puikių rezultatų. Vienas iš pavyzdžių yra Oxfor Micro Devices

A236 16-bit DSP procesorius. Jis turi keturis 16x16 bitų lygiagrečiai dirbančius daugybos įtaisus, o kiekvienas iš jų turi po 40 bitų sumos akumuliaciją, numatytą lygiagretaus jungimo galimybę. Sensory Rsc-300/364 kalbos atpažinimo įrenginiai, nors ir skirti balso atpažinimui neuroniniu tinklu, tačiau jų veikimas grįstas nuoseklia mikrokontrolerių architektūra.

Kita neuroprocesorių klasė yra neurospecifinės lygiagrečios architektūros neuroprocesoriai. Lyginant su nuosekliąja struktūra ši klasė yra patraukli savo skaičiavimo greičiu. Tai pasiekiama dideliu lygiagrečiai veikiančių skaičiavimo komponentų skaičiumi integrinės schemos viduje ir tai vadinama sistemos lygiagretumo laipsniu. Lygiagrečiai veikiančių komponentų skaičius sistemose svyruoja nuo 1 iki 10^6 ir daugiau lygiagrečiai dirbančių komponentų.



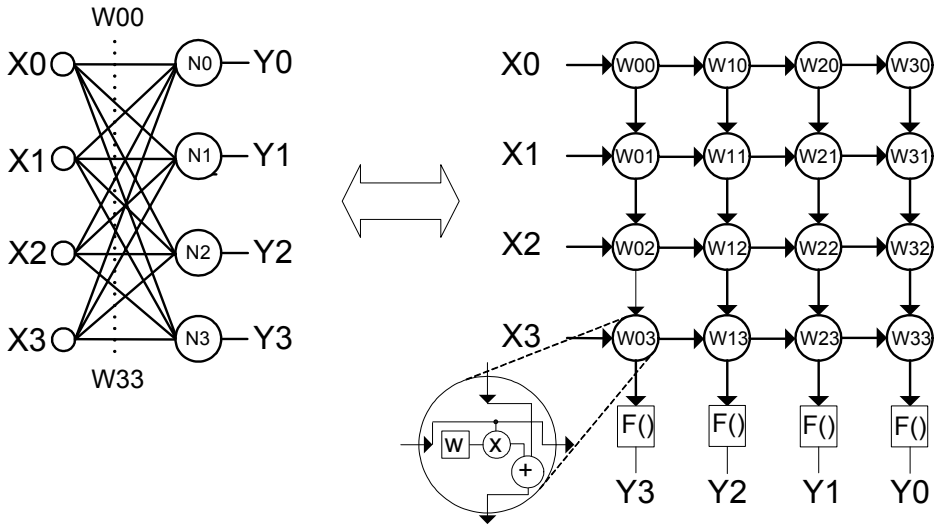
4 pav. Neuroprocesorių skirstymas pagal architektūrą

Kuo didesnis sistemos lygiagretumo laipsnis, tuo didesnis lusto plotas ir kaina. Dėl to NT aparatūrinės įrangos projektuotojai stengiasi sukurti kuo paprastesnius informacijos apdorojimo komponentus. Hibridinės struktūros pavyzdys būtų Heidelbergo universiteto sukurtas HAGEN (angl. *The Heidelberg AnaloG Evolvable Neural Network*) neuroprocesorius turi 32768 analogines sinapsių jungtis viename luste [11].

Neuroprocesorius pagal lygiagretumo laipsnį dar galima skirstyti į neuroninio tinklo sluoksnių, neuronų, sinapsių bei mišrų lygiagretinimą [9].

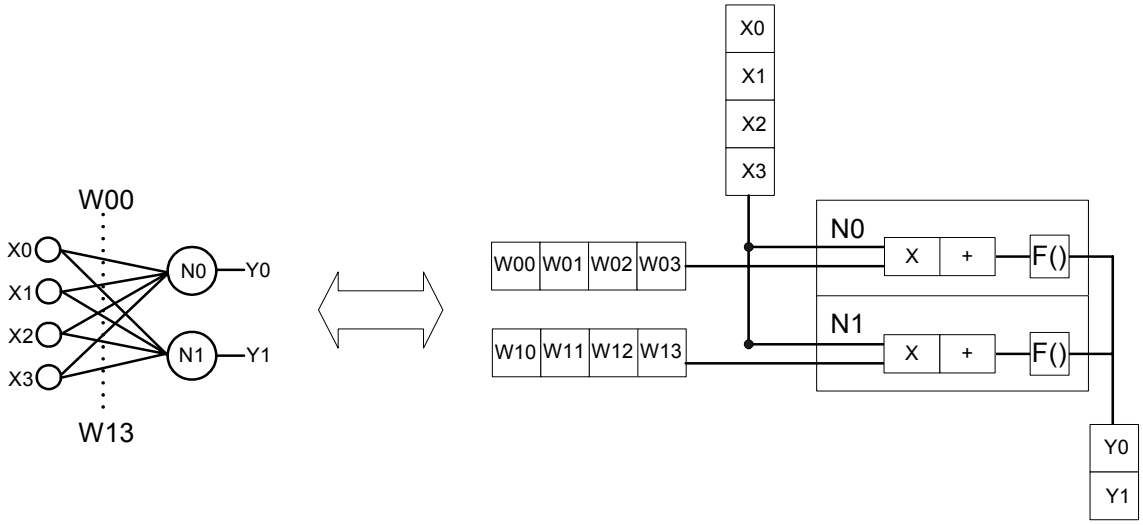
Sinapsių lygiagretinimas yra efektyviausias greičio atžvilgiu, nes neurono išėjimas paskaičiuojamas vienu CLK ciklu. Tačiau tokios architektūros neuroprocesorius gali talpinti mažus neuroninius tinklus, nes užimamas integrinės schemos plotas yra didelis, sunku dinamiškai keisti svorius mokymo metu. Panaudojus sinapsių lygiagretinimo architektūrą neuroprocesorius tampa sunkiai pritaikomas, nes neuroninis tinklas efektyviai skaičiuoja tik tokius tinklus, kurių sinapsių skaičius sutampa su neurokomponentų sinapsių skaičiumi.

Siekdami padidinti tokio tipo neuroprocesorių lankstumą projektuotojai stengiasi struktūrą daryti nuoseklesnę, keisdami svorių pakrovimą dinamišku kiekvienam neuronui atskirai. Kadangi nereikalingi itin tikslūs skaičiavimai neuroniniams tinklams, skaitmeniniai daugybos komponentai keičiami į analoginius, taip pasiekiamas mažesnis neurokomponento plotas ir padidinama skaičiavimo greitaveika.



5 pav. Sinapsių lygiagretinimo pavyzdys

Neuroprocesoriuose taip pat galima panaudoti neuronų skaičiavimo lygiagretinimą. Vienas iš būdų pateiktas 6 paveiksle. Sinapsių jungtys yra skaičiuojamos nuosekliai, tačiau neuronai tai daro lygiagrečiai.



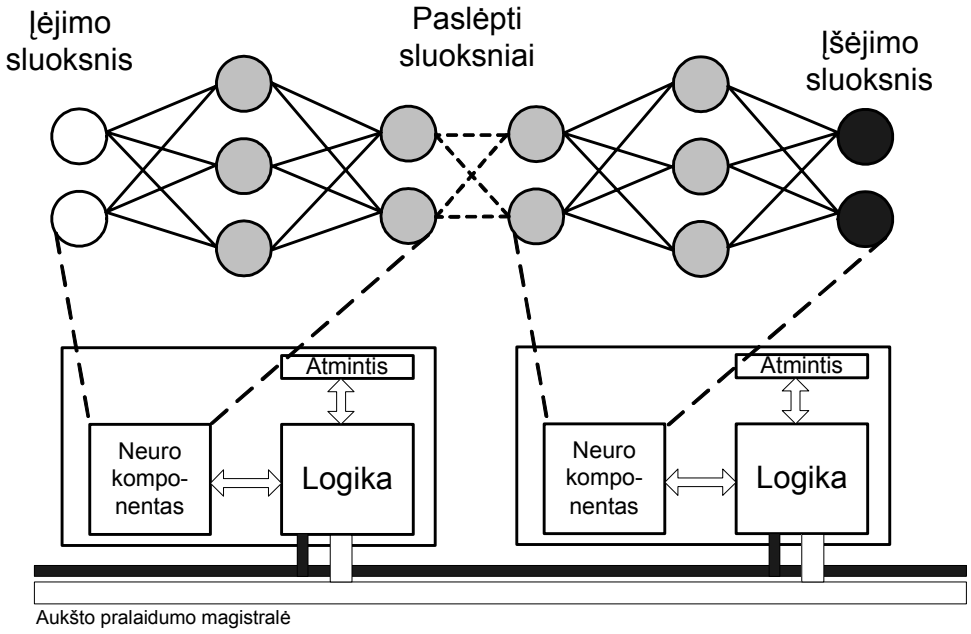
6 pav. Neuronų lygiagretinimo pavyzdys

Tokios architektūros privalumas yra kompromisas tarp sinapsinio ir sluoksninio tinklo lygiagretinimo. Vienam neuronui pakanka vieno daugybos įtaiso dėl to neurokomponento plotas yra pakankamai mažas, o greitaveika išlaikoma gana didelė, kad būtų pritaikoma daugeliui uždavinių spręsti.

Neuroninio tinklo sluoksnių lygiagretinimas leidžia jungti įvairius neurokomponentus į vieną struktūrą, galinčią apdoroti didesnius ir sudėtingesnius neuroninius tinklus. Neurokomponentais gali būti tiek klasikiniai nuosekliosios ar neurospecifinės lygiagrečiosios architektūros lustai, tiek įvairios skaičiavimo architektūros integrinės schemos viduje. Neurokomponentai į visumą jungiami bendra aukšto pralaidumo magistrale, su kuria jie jungiamas per valdymo logiką. Nuo magistralės struktūros priklauso neurokomponentų skaičius sistemoje, greitaveika, neuroninio tinklo struktūra. Sistemos greitis yra maksimalus, kai visi lygiagretūs komponentai atlieka darbą, o ne yra laukimo režime.

Paprastai aparatūrinėje įrangoje yra taikomos transliacinės (angl. *Broadcast bus*), linijinės (angl. *linear array*), žiedinės (angl. *systolic ring*), įvairaus tipo tinklinės (angl. *crossbar, bidimensional mesh*) magistralės [2].

Vienas iš žiedinės architektūros (angl. *Ring Array*) pavyzdžių yra ICSI (angl. *International Computer Science Institute, Berkeley, CA*) pagamintas neurokompiuteris, kurio pagrindiniai skaičiavimo komponentai yra standartiniai bendros paskirties procesoriai



7 pav. Neuroninio tinklo sluoksnių lygiagretumo pavyzdys

Neurokompiuterio sandara susideda iš 4-40 Texas Instruments TITMS320C30 slankaus kablelio DSP procesorių, sujungtų į žiedinę magistralę per Xilinx programuojamą logiką. Šis neuroprocesorius nuo 1990 m. buvo naudojamas balsui atpažinti [2].

Adaptive Solutions gaminamo neurokompiuterio CNAPS (angl. *Connected Network of Adaptive Processors*) yra pritaikyta transliacinės architektūros magistralė. Tokios magistralės privalumas yra tai,

kad sistemą galima greitai praplėsti pridedant naujų neuroprocesorių. Tačiau didelis neuroprocesorių skaičius sistemoje nėra efektyvus, nes visi neuroprocesoriai turi bendrauti bendra duomenų magistrale.

Žiedinės magistralės kaip ir transliacinės magistralės yra išplečiamos pridedant papildomų neurokomponentų, kiekvienas komponentas atstoja neurono vaidmenį, o tiksiklio ciklų skaičius neuroniniam tinklui apskaičiuoti lygus komponentų skaičiui. Žiedinės struktūros neuroprocesoriai gali skaičiuoti tiek tiesioginio, tiek grįžtamojo sklidimo neuroninius tinklus [10]. Tačiau atminties resursai išnaudojami neoptimaliai, nes kiekvienas neurokomponentas žiede turi turėti svorį kiekvienam iš žiedinės magistralės neurokomponentui.

Papildomų problemų neuroprocesorių projektavime atsiranda įgyvendinant sudėtingas perdavimo funkcijas tokias kaip logistinio sigmoido, hiperbolinio tangento ar radialinės bazės (2 pav.). Neuroprocesorių projektuotojai šia problemą sprendžia perdavimo funkcijas realizuodami paieškos lentelėmis(angl. *lookup tables*) arba aproksimuodami funkcijas tiesėmis. Tokiu būdu išvengdamas sudėtingų funkcijų skaičiavimas. Šiaip ar taip ne retai perdavimo funkcijų komponentai yra iškeliamos į lusto išorę, siekiant sumažinti lusto plotą bei padidinti lankstumą.

2.3. Projektavimo erdvės apžvalga

Tiek nuoseklios, tiek lygiagrečios struktūros aparatūrinė įranga turi savų pranašumų, kurie pastebimi plačiame taikymų spektre. Vienas pagrindinių nuoseklios aparatūrinės įrangos privalumų yra lankstumas, nes ji nesunkiai leidžia pakeisti programinę įrangą neliečiant aparatūrinės dalies. Lygiagrečios struktūros aparatūrinė įranga turi ženklius greičio apdorojimo parametrus. Tačiau keičiantis sistemos reikalavimams ar uždavinio sąlygoms, reikalingas brangus naujas aparatūrinės įrangos kūrimo bei diegimo procesas. Kompromisas tarp nuoseklių sistemų lankstumo ir lygiagrečių sistemų greičio yra lengvai modifikuojama integrinė grandinė – FPGA [13] [14]. Kadangi FPGA gali greitai keisti sklendžių matricą, ši technologija ypač tinkama integrinių schemų kūrimo stadijoje, nes leidžia nesunkai praktiškai patikrinti vieną ar kitą uždavinio sprendimo variantą.

2.3.1. FPGA įrenginiai

Yra daug FPGA struktūros variantų, tačiau visos šios grandinės susideda iš programuojamų loginių blokų ir juos jungiančio programuojamo tinklo [14]. Kai kure procesoriai, pavyzdžiui, Actel gamybos FPGA procesorius AX2000, be loginių blokų, turi ir atminties masyvų, Xilinx gaminamo Virtex II, Spartan klasių FPGA procesoriai turi konfigūruojamus daugybos įrenginius. Kai kurie iš jų turi integruotus mikroprocesorius.

Greitam neuroninio tinklo skaičiavimui reikalingi greiti bent 16x16 bitų daugybos įrenginiai bei greita daugybos įrenginio sąsaja su neuroninio tinklo svorių resursais. Konstruoti daugybos įrenginius iš standartinių FPGA loginių blokų yra neefektyvu. Tinkamai parinkus FPGA procesorių su integruotais aritmetiniais įrenginiais, galima gerokai sumažinti integrinei schemai reikalingos sklendžių matricos dydį. Taigi integruoti daugybos įtaisai ir atminties blokai FPGA įrenginyje yra privalomi siekiant sukurti greitaeigį neuroprocesorių.

Neuroprocesoriaus sukūrimas FPGA technologijoje su dideliu skaičiumi lygiagrečių neuroninių skaičiavimų yra sudėtingas, sunkiai įgyvendinamas procesas. Tam įtakos turi pakankamai aukšta FPGA procesorių kaina, be to, neuroninių tinklų skaičiavimui reikalingos daugybinės lygiagrečios daugybos operacijos ir atminties resursų poreikis yra sunkiai įgyvendinamas. FPGA gamintojai nuolat stengiasi pagaminti kiek galima didesnio sklendžių matricos dydžio įrenginius, turinčius ne vieną aritmetinį įrenginį. Pavyzdžiui, Xilinx sukurtas Virtex-4 klasės FPGA procesorius XC4VFX140 turi 552 18Kb atminties blokus, 192 18x18 bitų daugybos įtaisus, 2 integruotus PowerPC klasės procesorius gali atlikti našius neuroninius skaičiavimus, integruoti valdikliai leidžia taikyti sudėtingesnius mokymo algoritmus.

Xilinx įmonės Virtex II klasės XC2V40 FPGA procesorius turi 4x18bitų atminties blokus, keturis 18x18 bitų fiksuoto kablelio daugybos įtaisus, o sklendžių matrica gali sutalpinti iki 40K ventilių integrinę schemą. Šio FPGA procesorius pakanka, kuriant bei tiriant mažos integrinės schemos ploto neuroninio procesoriaus modelį.

2.3.2. Aparatūrinės įrangos projektavimo kalbos

Pagrindinės aparatūrinei įrangai projektuoti ir modeliuoti skirtos programavimo kalbos yra Verilog, VHDL, SystemC, HandelC. Seniausios ir populiariausios iš jų yra Verilog ir VHDL [17]. Pagal struktūrų aprašymą šios dvi klabos yra labai panašios. Abi gali aprašyti įvykių abstrakcijas ir hierarchines struktūras. Aprašomos struktūros susideda iš blokų, kuriuos gali sudaryti ir kiti blokai, ventiliai, galimybė procedūromis aprašyti lygiagrečius procesus, sąsajos tarp modulių nusakomos vienareikšmiškai. Tiesa Verilog programavimo kalba yra kompaktiškesnė, užrašomas kodas trumpesnis ir aiškesnis, tačiau negalima kurti tipo struktūrų bei vietinių kintamųjų. VHDL yra gerokai lankstesnė leidžianti aprašyti platesnius modeliavimo bei sintezavimo uždavinių sprendimo variantus.

Laikas parodė, kad Verilog ir VHDL programavimo kalbomis, gebančiomis aprašyti sudėtingus elgsenos ir registrų perdavimo lygmens uždavinių sprendimo variantus žemame ir vidutiniame abstrakcijos lygyje, sunku aprašyti sudėtingas sistemas turinčias aukštą abstrakcijos lygį. Sunku modeliuoti aparatūrinę įrangą turinčią procesorinius modulius, įterptinę programinę įrangą ir kitus

sudėtingus IP (angl. *Intellectual property*– intelektuali nuosavybė) blokus [18]. SystemC ir HandelC yra populiariausios C++ pagrindu sukurtos programavimo kalbos. Jas naudojant nesunkiai modeliuojami tiek aukšto, tiek žemo abstrakcijos lygį turintys komponentai. Galima aprašyti elgsenos bei registrų perdavimo struktūrinius komponentus. HandelC sukurtas pagal ANSI-C standartą, System-C tai C++ klasių biblioteka aprašanti aparatūrinės įrangos specifiką C++ modeliams.

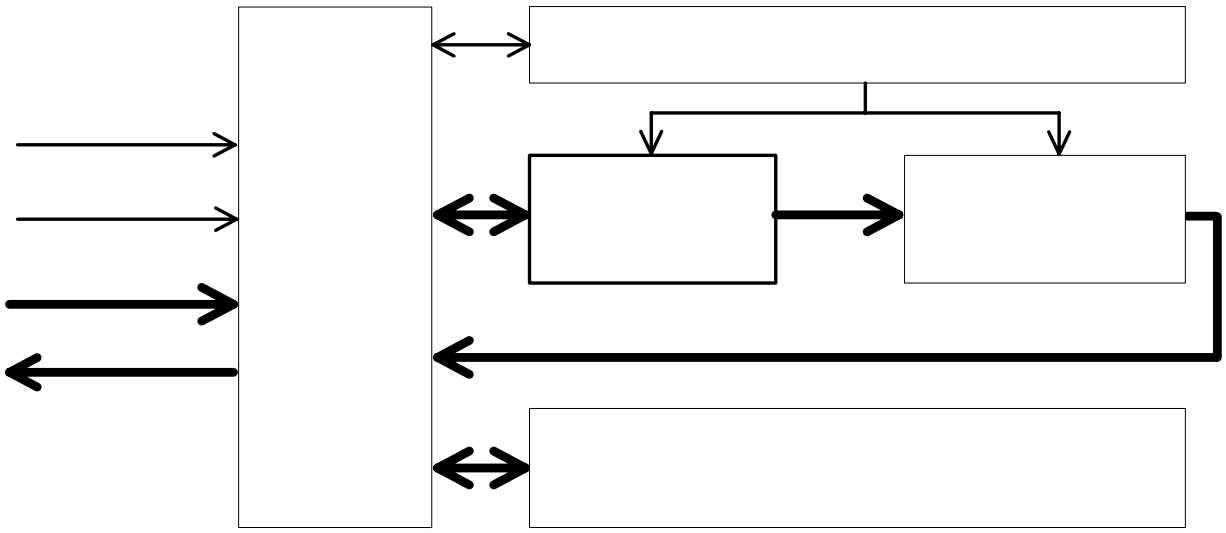
Neuroprocesoriai paprastai nėra autonominiai įrenginiai. Paprastai jie būna sujungti su kitais neuroprocesoriais, valdančiais moduliais, sensoriais ar valdikliais. System-C kalba leidžia modeliuoti tokius tarpusavio įrenginius it informacijos srautus tarp jų. Tuo tarpu Verilog ar VHDL tinkamos neuroprocesoriaus vidinių struktūrų analizei. Kai kurie sintezės įrankiai, pavyzdžiui, CoCentric System Studio gali atlikti konvertavimo iš System-C į Verilog ar VHDL programavimo kalbas. Todėl System-C patogu naudoti kaip modeliavimo kalbą, o Verilog ir VHDL analizei bei sintezei.

3. Projektinė dalis

Šioje dalyje aprašoma neuroprocesoriaus prototipo kūrimo iš aukšto abstrakcijos lygmens į FPGA technologiją realizacijos būdas. Pasiūlomas apibendrintas neuroprocesoriaus modelis. Remiantis juo suformuojamas nuoseklios struktūros matematinis neuroprocesoriaus modelis, tiriamos jo charakteristikos. Matematinis modelis realizuojamas SystemC programavimo kalba.

3.1. Apibendrintas neuroprocesoriaus modelis

Paskutiniais dešimtmečiais sukurta daug neuroprocesorių modelių, besiskiriančių realizuotomis architektūromis, veikimo greičiu lygiagretumo laipsniu ir kt. Šiame skyriuje pasiūlytas apibendrintas neuroprocesoriaus modelis atspindi pagrindines neurospecifinių procesorių bruožus (8 pav.).



8 pav. Apibendrintas neuroprocesoriaus modelis

Išskiriamos penkios pagrindinės dalys: neuroprocesoriaus širdis – daugybos ir sumavimo blokas, perdavimo funkcijos blokas, sąsajos blokas, atmintis ir valdymo įrenginys. Daugybos ir sumavimo įrenginys visada būna neuroprocesoriaus viduje. Atmintyje paprastai saugomi sinapsių svoriai, neuronų išėjimo reikšmės, gali būti saugoma ir neuroninio tinklo struktūrą. Atminties struktūros gali būti išbarstytos visame neuroprocesoriuje kiekvienam neurokomponentui atskirai arba vientisos, priklausančios visam neuroprocesoriui. Taip pat atmintis gali būti tiek neuroprocesoriaus viduje, tiek išorėje. Įrenginiai, atliekantys perdavimo funkcijų vaidmenį, paprastai būna neuroprocesoriaus viduje, tačiau kartais būna iškelti į išore, siekiant sumažinti lusto plotą. Sąsajos blokas apibūdina

Neuroninis tinklo struktura.

Valdymo parametrai

Sąsajos

neuroprocesoriaus architektūrą. Nuo jo priklauso kokios architektūros magistralėmis bus sujungti neurokomponentai luste, kaip bus interpretuojama atmintis, išorinių įrenginių bendravimo principas.

3.2. Uždavinio analizė

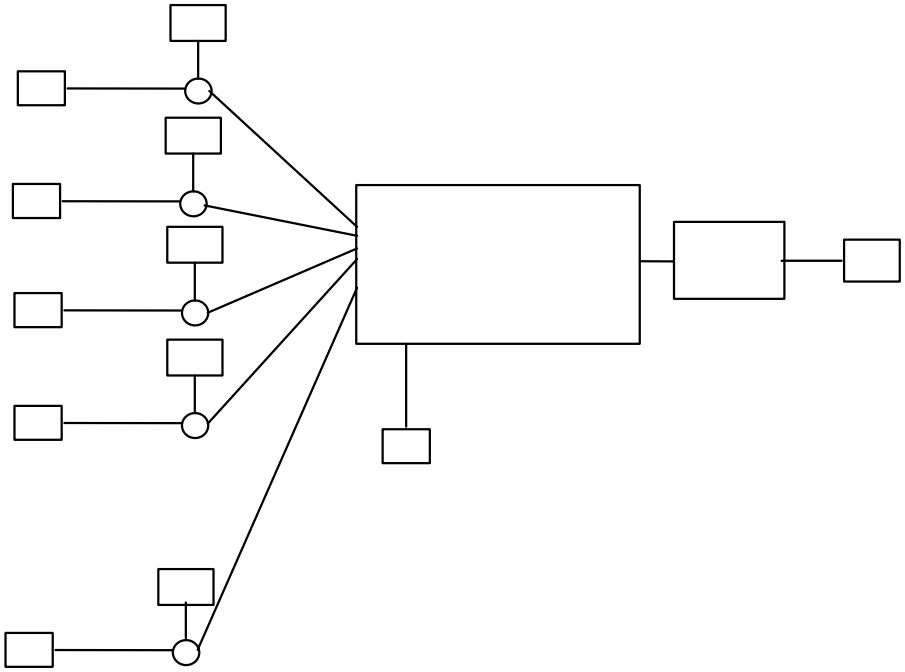
Projektuojant neuroprocesoriaus modelį būtina sudaryti pagrindinius vertinimo kriterijus. Kuriamas neuroprocesoriaus modelis turi turėti:

- architektūrą, galinčia turėti įvairų lygiagrečiai veikiančių neurokomponentų skaičių;
- galimybę keisti neuroninio tinklo struktūrą darbo rėžime;
- bent 10 bitų duomenų plotį;
- grįžtamojo ryšio neuroninių tinklų palaikymą;
- bent pora perdavimo funkcijų;
- dinamiškai keičiamą neuroninio tinklo dydį.

Nuo pasirinktos neuroprocesoriaus architektūros dažnai priklauso maksimalus ir minimalus lygiagrečiai veikiančių neurokomponentų skaičius. Tai svarbu, nes pritaikant neuroprocesorių skirtingiems FPGA įrenginiams neurokomponentų skaičius kinta. Pritaikant įvairius mokymo algoritmus reikalingas nuolatinis neuroninio tinklo struktūros keitimas. Kartais pasikeitus aplinkos parametrams taip pat reikalingas neuroninio tinklo struktūros keitimas, nenutraukiantis neuroprocesoriaus darbo. Rinkoje yra įvairaus duomenų pločio neuroninių procesorių nuo kelių iki 16 bitų ir daugiau. Dažnai dėl mažo neuroprocesoriaus duomenų pločio neuroprocesoriai, nėra pritaikomi uždaviniams kuriems reikalingas didelis tikslumas. Dėl to bent 10 bitų plotis yra būtinas. Grįžtamojo ryšio neuroniniai tinklai leidžia įgyvendinti dinamiškas sistemas, o keletas perdavimo funkcijų tik padidina sistemos lankstumą. Neuroprocesoriaus lankstumą didina ir tai, kad jis gali apdoroti negriežtai nustatyto dydžio neuroninius tinklus, o dinamiškus- tiek mažus tiek didelius.

3.3. Matematinis neuroninio tinklo modelis

Neurono matematinė išraiška yra pakankamai nesudėtinga (9 pav.). Neurono įėjimai $X_{n,m}$, kuriais gali būti kitų ar to paties neurono išėjimas bei neuroniniam tinklui nepriklausantys kintamieji, yra dauginami iš svorių $S_{n,m}$. Skaičių $X_{n,m}$ ir $S_{n,m}$ sandauga formuoja neuronų sinapses. Sinapsės biologiniuose neuronuose būna signalą stiprinančios arba slopinančios, dėl to svoriai matematiniame neurono modelyje taip pat turi pasižymėti šiomis savybėmis, tai yra $S \in \mathbf{R}$. Kai $S \in (-1;1)$, tai sinapsė signalą silpnina, o visi likę realūs skaičiai signalą stiprina. Sinapsių formuojamos reikšmės yra sumuojamos, o galutinę sumą koreguoja aktyvacijos slenkstis $B_{n,}$. Perdavimo funkcija $F(a_n)$ priklausomai nuo sukaupto a_n lygmens praleičia signalą toliau arba ne. Dažniausiai sutinkamos perdavimo funkcijos pavaizduotos 2 paveiksle.

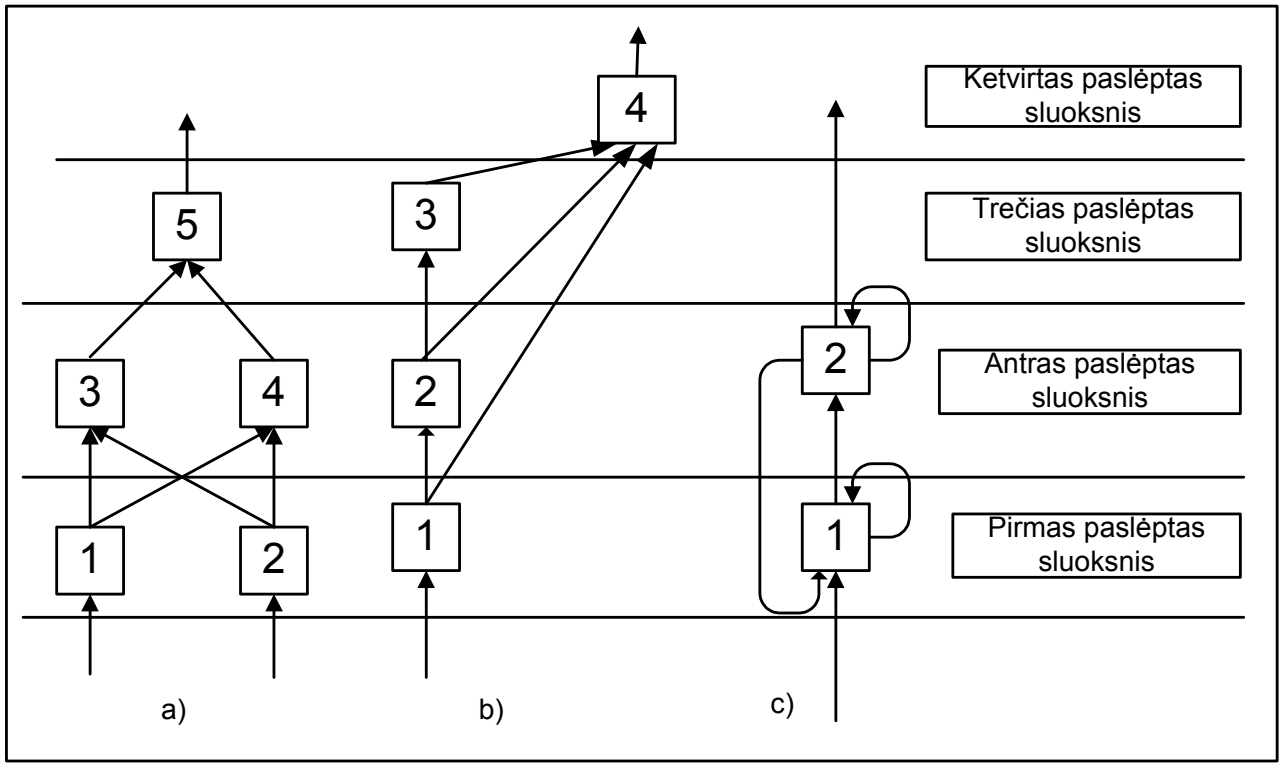


9 pav. Matematinis neurono modelis

$$Y_n = F_n(\sum_{m=0}^{M_n} X_{n,m} \bullet S_{n,m} + B_n) \qquad [1]$$

Matematinėje neurono modelio formulėje [1] M-neurono sinapsių skaičius, n - neurono eilės numeris, X-neurono įėjimai, S- neurono svoriai, B- neurono aktyvacijos slenkstis, F-neurono perdavimo funkcija.

Neuronai jungdamiesi vieni su kitais formuoja sudėtingus neuroninius tinklus, dėl to matematinė jų išraiška tampa sudėtinga. Trijų tipų neuroninių tinklų pavyzdžiai pavaizduoti 10 paveiksle. Jei neuronus sunumeruotume eilės tvarka nuo pirmo iki paskutinio sluoksnio, tada neuroninį tinklą galima užrašyti formule (2 form.).



10 pav. Neuroninio tinklo skaičiavimo eiliškumas a) tiesioginio sklidimo sluoksninis neuroninis tinklas b)tiesioginio sklidimo neuroninis tinklas c) grįžtamojo ryšio neuroninis tinklas

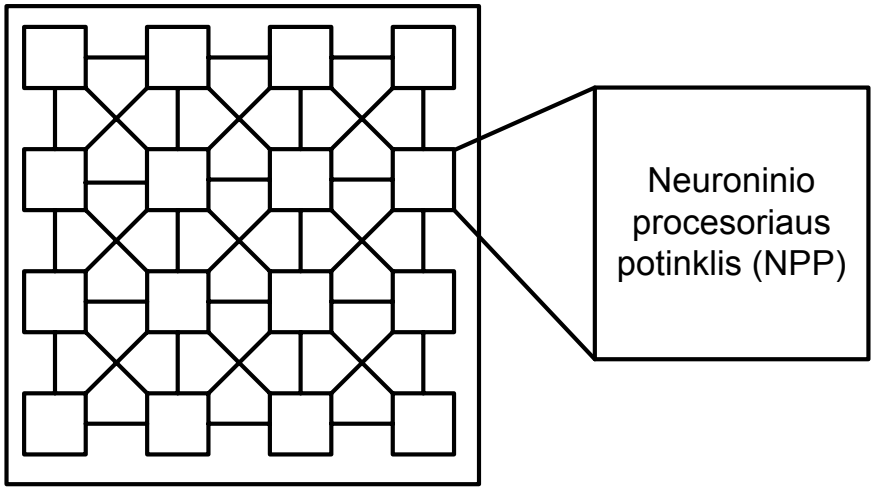
$$Y(n) = F_n\left(\sum_{m=0}^{M_n} Y(A_{n,m}) \bullet S_{n,m} + B_n\right) \qquad [2]$$

Čia n=[1;N]; N-neuronų skaičius; M-neurono sinapsių skaičius; Y-neurono išėjimo reikšmė arba neuroninio tinklo įėjimas; A-neuronų adresų sąrašas; S - neuronų svorių sąrašas, B-neuronų sužadinimo slenkstis, F-neurono perdavimo funkcija.

Visas neuroninio tinklo ciklas praeis, kai bus apskaičiuotos visos Y(n) reikšmės intervale [1;N]. Neuroninio tinklo sluoksniai skaičiuojami nuo pirmo sluoksnio paeiliui iki paskutinio. Neuroninius tinklus pavaizduotus 10 paveiksle galima aprašyti 2 formule.

3.4. Neurokomponentų sąsaja

Nuo pasirinktos neurokomponentų sąsajos priklauso tolimesni neuroninio tinklo projektavimo etapai. Panašiausios į biologinius tinklus yra tinklinės magistralių architektūros (11 pav.). Suskirsčius neuroninį tinklą į daugybę mažų potinklių, kurių kiekvienas turėtų ryšius su arčiausiai esančiais „kaimynais“, galima pasiekti aukštą lygiagretumo laipsnį. Be to, keičiant NPP skaičių tinkle galima neuroprocesorių pritaikyti skirtingo dydžio FPGA įrenginiams.



11 pav. Neuroninis tinklas suskirstytas į potinklius

Tinklinė neuroprocesoriaus architektūra pasižymėtų:

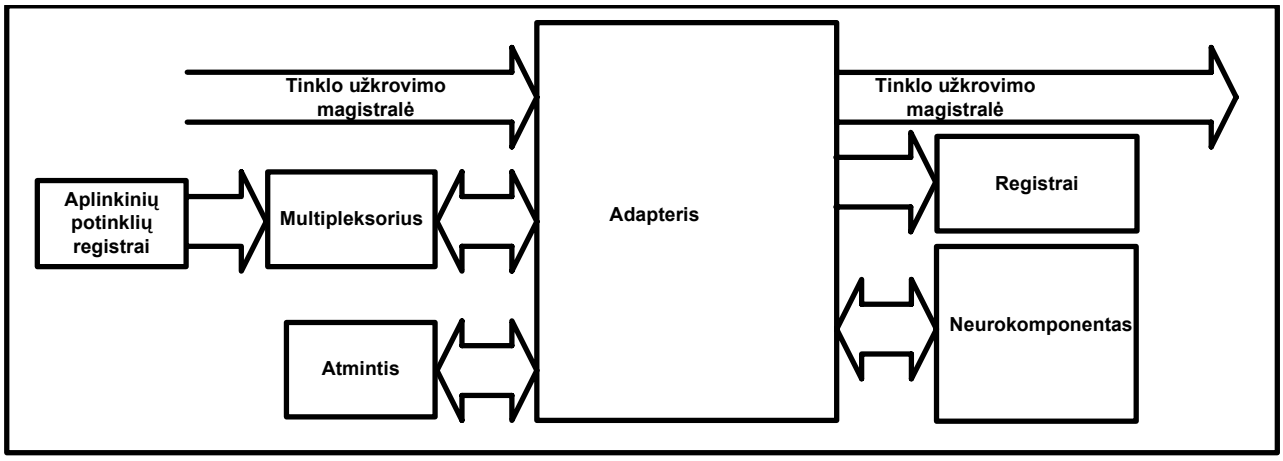
- struktūrų vienodumu ir paprastumu;
- neribojamu lygiagretumo laipsniu;
- tiesioginio sklidimo ir grįžtamojo ryšio tinklų palaikymu;
- panašumu į biologinių neuroninių tinklų struktūras;
- patikimumu - sugedus vienam potinkliui jo darbą galėtų kompensuoti šalia esantis.

Tinklinės magistralės informaciją gali perduoti tiek nuosekliai, tiek lygiagrečiai. Informacija perduodant nuosekliomis magistralėmis, galima sudaryti daugiau ryšių tarp neuroprocesoriaus potinklių, tačiau lygiagrečios magistralės yra pralaidesnės informacijai. Daugybos komponentai, esantys FPGA įrenginiuose, operacijas atlieka vienu tiksiklio ciklu, dėl to taikyti nuoseklias magistrales duomenų mainams yra netikslinga.

NP sąsaja turi užtikrini šių problemų sprendimus:

- neuroninio tinklo pakrovimą į procesoriaus ląsteles;
- neuroninio tinklo duomenų pateikimą neuroniniam potinkliui.

NPP(neuroninio procesoriaus potinklio) duomenys, tokie kaip neuronų reikšmės, neuronų svoriai ir adresai turi būti saugomi atminties struktūrose. Kadangi šių duomenų yra daug (>1KB), pravartu naudoti RAM tipo atmintis. Neuroninio tinklo pakrovimą į atmintį galima realizuoti naudojant adapterį tarp atminties struktūrų ir neurokomponento (12 pav.). Atskyrus atmintį nuo NP, supaprastėja sistemos struktūra, nes visos neurokomponento atliekamos operacijos būtų tik su adapterio formuojama virtualia atmintimi. Adapteris virtualią atmintį formuotų iš atminties bloko, kuriame saugoma neuroninio tinklo struktūra, potinklio išėjimo registrų ir aplinkinių potinklių išėjimo registrų. Šiuo būdu realizavus neuroprocesoriaus potinklį, sistema tampa lanksti, nes galima keisti adapterio struktūrą, atminties dydį, išėjimo registrų skaičių, nekeičiant pagrindinio neurokomponento logikos.



12 pav. Blokinė NP potinklio schema

Apibendrinus sąsajos neurokomponentus, neuroprocesorius tampa pritaikomas įvairaus dydžio FPGA įrenginiams.

3.5. Neurokomponentas

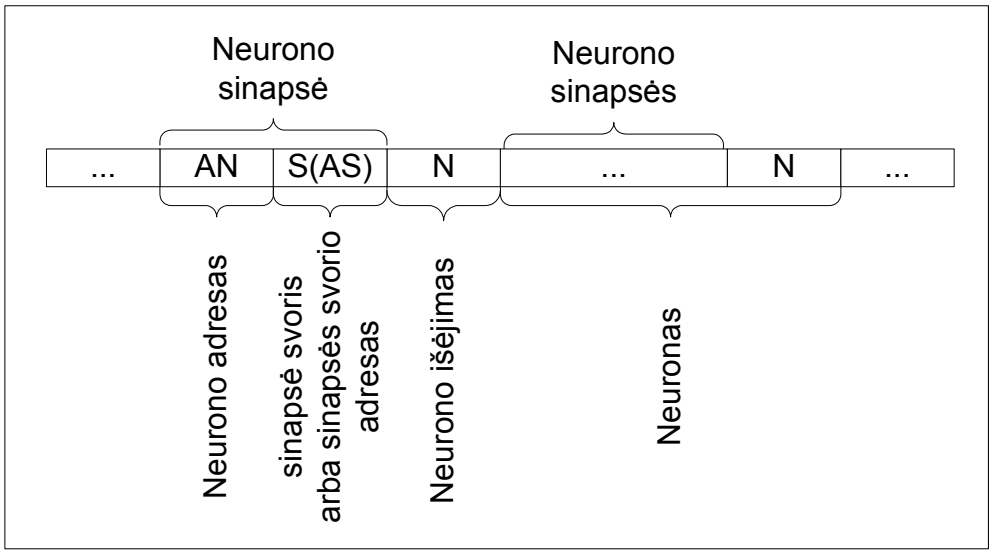
Šioje dalyje nagrinėjamas nuoseklaus neurokomponento realizavimo būdas NPP bloke.

3.5.1. Neuroninio tinklo struktūra

Analizuojant neuroninį tinklą galima išskirti tokius duomenų sąrašus (2 form):

- neuronų išėjimų sąrašą Y(n);
- neuronų svorių sąrašus $S_{n,m}$;
- neuronų adresų sąrašus $A_{n,m}$, kurie nurodo su kuo neuronas n jungiasi ;
- neuronų aktyvacijos slenksčio sąrašą B_n.

Duomenų struktūros, gali būti saugomos viename arba keliuose atminties blokuose. Paprastai atminties RAM blokų FPGA įrenginiai turi tiek pat kiek ir daugybos įtaisų, todėl neuroninio tinklo struktūrom saugoti tikslinga naudoti vieną RAM bloką.



13 pav. Neuroninio tinklo struktūros elementų išsidėstymo eiliškumas atminties bloke

Kadangi neuroninius skaičiavimus neurokomponente atliks tik vienas daugybos įtaisas, skaičiavimai yra nuoseklūs. Dėl šios priežasties svarbu numatyti NPP skaičiavimo eiliškumo seką. Siekiant supaprastinti uždavinį, neuronų sužadinimo slenksčiai bus traktuojami kaip paprastos sinapsės, kurių svoriai signalo stiprumo nekeičia. Tuo būdu galima išvengti perteklinių grandinių integrinėje schemoje, tačiau neapsieinama be papildomų laiko išteklių.

Neuroninio tinklo struktūrą galima išdėstyti 13 paveiksle pateiktu eiliškumu. Neuronai atmintyje išdėstomi pagal skaičiavimo eiliškumą. Nusakant neuroną, pirmiausia eina sinapsės, po to neurono

ląstelės. Neurono sinapsę sudaro du atminties komponentai, užimdami dvi atminties ląsteles. Viena iš jų yra svoris arba nuoroda į svorį, kita neurono adresas, su kuriuo sudaryta sinapsė. Čia pateiktame modelyje pirmoji iš dviejų atminties ląstelių nusako adresą, o antra svorį. Svorio ląstelė nebūtinai turi būti skaičius, jis taip pat gali būti adresas neurono, kurio reikšmė traktuojama kaip svoris. Tuo būdu galima sukurti neuroninį tinklą, dinamiškai keičiamais svoriais, to paties neuroninio tinklo.

1 lent. Neuroninio tinklo duomenų struktūra atminties ląstelėje

Saugoma neurono struktūra	**Duomenys 12 bitų**	**Vėliavėlės 4 bitai**			
		Identifikacija		**Papildomos**	
Neurono adresas	XXXXXXXXXXXX	0	0	X	1- į pradžią
Svoris	XXXXXXXXXXXX	1	0	X	
Svorio adresas	XXXXXXXXXXXX	0	0	X	
Neuronas	XXXXXXXXXXXX	1	0	Perdavimo funkcija 1-tiesinė 0-tiesinė su ribojimu	
Išėjimo adresas	XXXXXXXXXXXX	0	1	X	

Kad NPP „suprastų“ kokios neuroninio tinklo ląstelės yra saugomos atmintyje, kiekvienas atminties masyvo elementas turi turėti vėliavėles, nusakančias kokia neurono dalis juose saugoma (1 lent.). Neurono struktūrinės dalies identifikavimui atminties ląstelėje užtenka 2 bitų. Dar du bitai reikalingi neurono perdavimo funkcijai nurodyti ir neuroninio tinklo struktūros pabaigos identifikacijai. Išviso vėliavėlėms reikia 4 bitų, likusius 12 bitų paliekant atminties neuroninio tinklo reikšmėm saugoti: svoriams, adresams, neuronų išėjimams.

3.5.2. Daugybos ir sumavimo blokas

Neurono sinapsės gali būti tiek stiprinančios signalą, tiek ir slopinančios. Kad sinapsės turėtų slopinantį poveikį, svorio modulis turi būti mažesnis už vienetą. Aparatūrinėje įrangoje duomenys yra sveiki skaičiai, tačiau mažesnius už vienetą svorius galima gauti įvedus papildomą dalybos veiksmą po kiekvienai daugybos operacijos (3 form.).

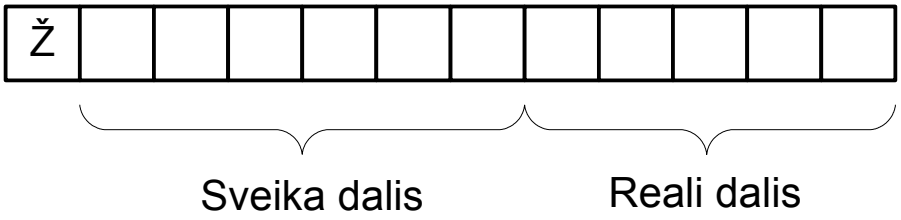
$$Y(n) = F(\sum_{m=0}^{M_n} Y(\frac{A_{n,m} \bullet S_{n,m}}{D} + B_n)) \qquad [3]$$

Aparatūrinėje įrangoje dalikliui D pravartu naudoti 2^x skaičių, tokių atveju apsieinama tik su poslinkio operacija.

2 lent. 11 bitų svorio intervalai keičiantis dalikliui

Daliklis	Svoris
4	0.250-512
8	0.125-256
16	0,062-128
32	0,031-64
64	0,016-32
128	0,008-16
256	0,004-8

Nuo daliklio priklauso svorio intervalas (2 lent.). Panaudojus 2^8 skaičių, sveikai daliai lieka tik 2^3 maksimalus galimas svoris. Dėl to, kad sinapsės būtų vienodai slopinančios ar stiprinančios, skilčių skaičių pravartu naudoti vienodą tiek sveikai tiek realiai daliai.



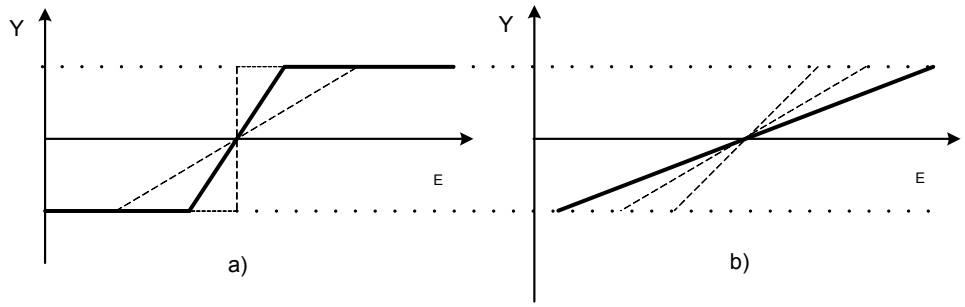
14 pav. Svorio duomens dedamosios

Tai yra, jei naudojamas duomenų plotis yra 12 bitų, tai šešios skiltys paliekamos sveikiems skaičiams, o penkios realiai daliai vaizduoti. Daliklis šiuo atveju bus 2^5=32.

Maksimalus neurono sinapsių skaičius neuroprocesoriuje priklauso nuo akumuliacinio registro pločio ir nuo sinapsinės sandaugos užimamų bitų pločio. Esant 12bitų pločio duomenims, 24bitų pločio akumuliaciniam registrui ir dalikliui lygiam 2^5 galima naudoti 127 sinapses vienam neuronui.

3.5.3. Perdavimo funkcijos

Paprasčiausios perdavimo funkcijos, kurioms nereikia sudėtingų matematinių skaičiavimų, yra šios: tiesinė, tiesinė ribojanti ir šuolinė funkcija. Neuroniniame procesoriuje panaudojus tiesinę ir tiesinę ribojančią funkciją, neuronų svoriais galima keisti funkcijos nuožulnumą. Panaudojus didesnius svorius neuronui su tiesine ribojančia funkcijai, gaunamas neuronas su šuolinė funkcija (15 pav.).



15 pav. Neuroninio tinklo perdavimo funkcijos

Perdavimo funkcijos aukščiausią ir žemiausią tašką visada riboja pasirinktas duomenų plotis. Funkcijos išėjimas visada turi tilpti į pasirinktą duomenų plotį. 4 formulėje min ir max atitinkamai žemiausias ir aukščiausias funkcijos taškas. Tiesinės funkcijos (5 form.) kintamuoju d parinkus 2^x skaičių, apsieinama tik su poslinkio operacija.

$$Y_n = F_0(a_n) = \begin{cases} \max & ,kai \quad a_n > \max; \\ a_n & ,kai \quad \min <= a_n <= \max; \\ \min & ,kai \quad a_n < \min \end{cases} \qquad [4]$$

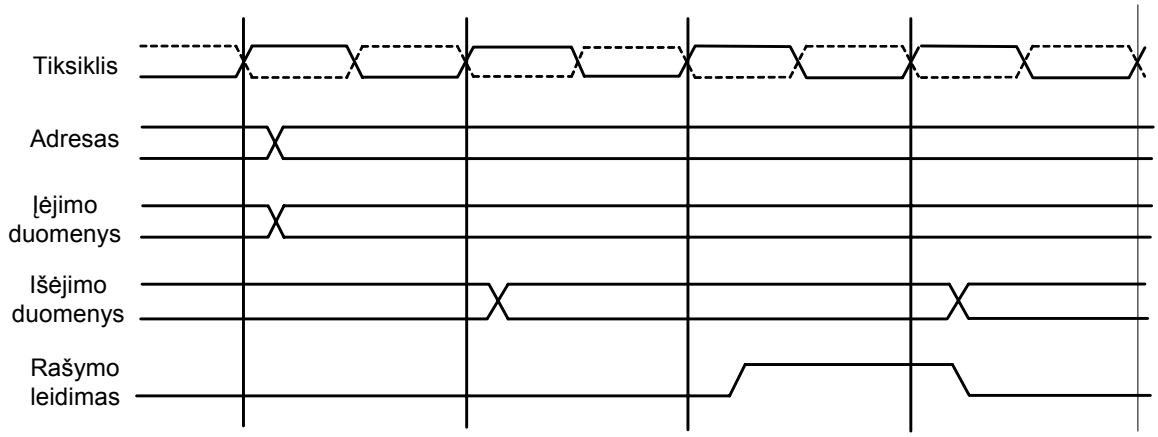
$$Y_n = F_1(a_n) = \frac{a_n}{d} \qquad [5]$$

Tiesinės funkcijos konstanta d turi būti parenkama tokia, kad į pasirinktą neurono išėjimo saugojimo registro plotį Y_n visada tilptų akumuliacinio registro skaičius. Kai akumuliacinio registro plotis yra 24 bitai, o pasirinktas neurono išėjimo plotis yra 12 bitų, tai d turi būti lygus 2^{12} skaičiui. Šiuo atveju tiesinė funkcija tampa dalybos operacija.

Neuroprocesoriuje naudojant tiesinę ir tiesinę ribojančią perdavimo funkciją sutaupomas integrinės schemos plotas, nes nereikalingi sudėtingi matematiniai skaičiavimai.

3.5.4. Neurokomponento skaičiavimo ciklų analizė

Nuoseklios struktūros neuroprocesoriaus didžiąją dalį skaičiavimų laiko sudaro informacijos nuskaitymas ir rašymas į atmintį. Vienos ląstelės nuskaitymui reikia sugaišti du tiksiklio ciklus, taigi nuskaityti rodyklei ir rodyklės rodomai reikšmei sugaištama dvigubai daugiau laiko (3 lent.).



16 pav. Atminties bloko RAM (angl. *random access memory*) laikinės diagramos

Neuroprocesorius daugybos ir sumavimo veiksmus gali atlikti informacijos skaitymo iš atminties momentais, taigi neuroninio procesoriaus skaičiavimo greitis priklauso tik nuo atminties skaitymo ciklų ir greičio (3 lent.).

3 lent. Atminties (RAM) ir daugybos įtaisų skaičiavimo ciklų pasiskirstymas laike

Apdorojamas neuronas	Atminties ląstelės tipas	Tiksiklio ciklų skaičius operacijai	Daugybos įtaiso, sumatoriaus, perd. funkc.aktyvumas
N_1	$A_{1,1}$	2	
	$NA_{1,1}$	2	
	$S_{1,1}$	2	
	$A_{1,2}$	2	*,+
	$NA_{1,2}$	2	
	$S_{1,2}$	2	
	N_1	2	*,+,F
	NS_1	1	
N_2	$A_{2,1}$	2	
	$NA_{2,1}$	2	
	$S_{2,1}$	2	
	N_2	2	*,+,F
	NS_2	1	
	O_2	2	
	NS_2	1	

Lentelėje $A_{n,m}$ –skaitomas sinapsės neurono adresas; $NA_{n,m}$- skaitomas neuronas adresu $A_{n.m}$, su kuriuo

sudaromą sinapsė, $S_{n,m}$- skaitomas sinapsės svoris, N_n- skaitomi neurono parametrai; NS_n –išsaugomas neurono išėjimas; O_n-išsaugomas neurono išėjimas potinklio išėjimo registre.

Siekiant nustatyti neuroprocesoriaus modelio skaičiavimo greitį, 3 lentelės pagrindu sudaroma neuroninio procesoriaus skaičiavimo ciklo trukmės formulė (10 form.).

$T_A=4\Delta t$ [6]

$T_S=2\Delta t$ [7]

$T_{NS}=3\Delta t$ [8]

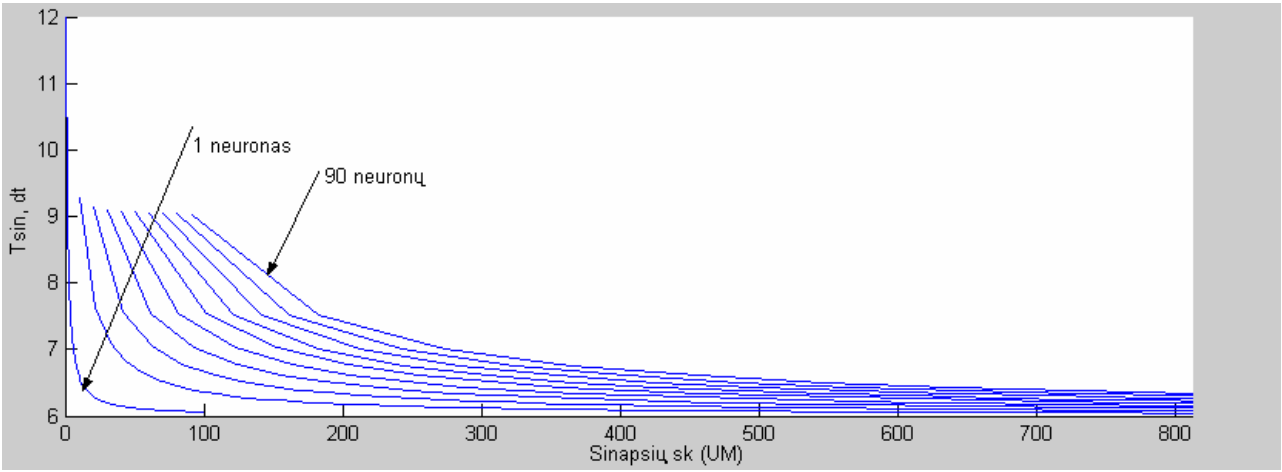
$T_n = M_n(T_A+T_S)+T_{NS}= 6M_n\Delta t+3\Delta t$ [9]

$$T_{net} = \sum_{n=0}^{N-1} T_n + N' \bullet T_{NS} = \sum_{n=0}^{N-1} (6M_n\Delta t + 3\Delta t) + 3N' \bullet \Delta t \quad [10]$$

$$T_{sin} = \frac{T_{net}}{\sum_{n=0}^{N-1} M_n} \quad [11]$$

kai T_A-adreso skaitymo trukmė; T_S-svorio skaitymo trukmė; T_{NS}-neurono nuskaitymo ir išsaugojimo trukmė; N'-Išėjimų skaičius; T_{sin}-sinapsės apskaičiavimo trukmė.

Tiksiklių ciklų skaičius vienam sinapsės apskaičiavimui vertinamas 17 paveiksle, esant skirtingam neuroninio tinklo dydžiui. Grafikas gaunamas panaudojus 11 formulę, kai M visiems neuronams yra lygus ir priklauso intervalui [1;100], neuronų skaičius neuroniniame tinkle priklauso aibei [1,10,20,30,...,90].

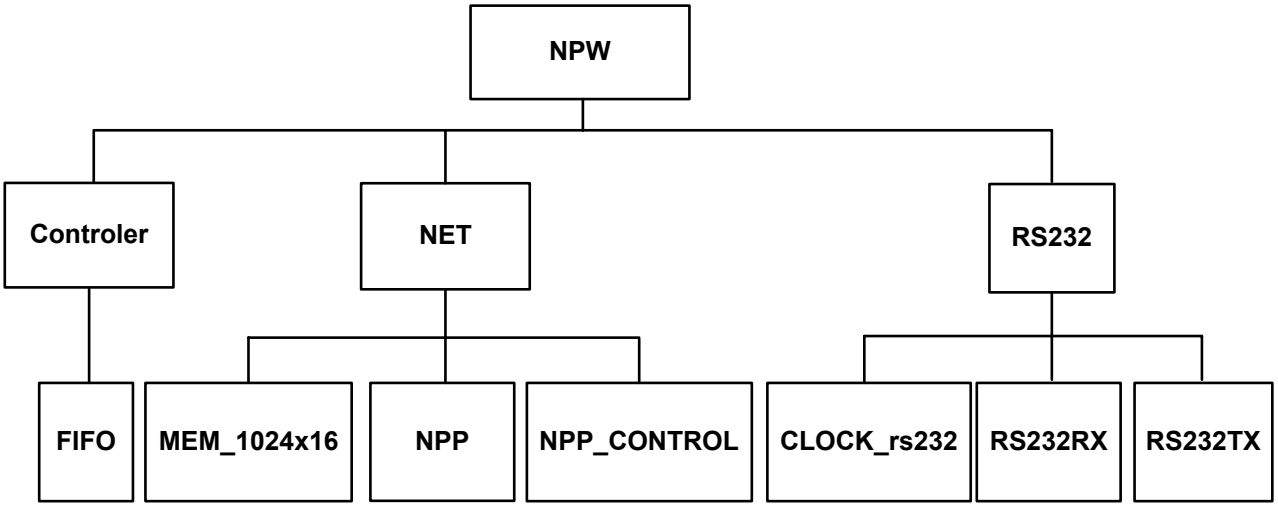


17 pav. Tiksiklio ciklų skaičius vienai sinapsei, esant skirtingam skaičiui neuronų neuroniniame tinkle

Matyti, kad skaičiavimo efektyvumas priklauso nuo neuroninio tinklo struktūros. Kuo daugiau neuronų neuroniniame tinkle, tuo skaičiavimo efektyvumas mažesnis.

3.6. SystemC realizacija

Neuroninis procesoriaus prototipas realizuojamas SystemC programavimo kalba. Suprogramuota registrų perdavimo lygyje (angl. RTL). Toks stilius buvo pasirinktas dėl to, kad yra labai svarbūs tokie modelio parametrai kaip greitis ir užimamas integrinės schemos plotas, o elgsenos lygmens programavimo stilius šiomis savybėmis nepasižymi. Sukurtas SystemC neuroprocesorius prototipas paprastumo dėlei turi vieną potinklį.

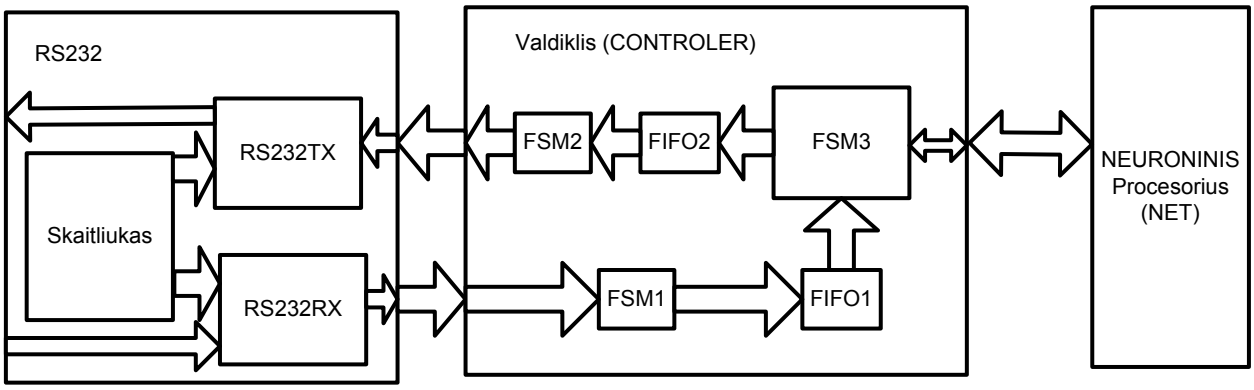


18 pav. Struktūrinė projekto schema

Sistemą SystemC kalboje sudaro neuroprocesorius (NET), RS232 įrenginys ir valdiklis (Controler). RS232 modulis atlieka RS232 protokolo kodavimą. Dėl RS232 protokolo žodžio ilgio (8 bitai.) ir neuroprocesoriaus duomenų ilgis (16 bitų) nesutapimo, reikalingas informacijos transformavimas. Tai atlieka "Controler" modulis, taip pat jis atlieka neuroprocesoriaus valdymą.

Panaudojus atvirkštinio dėklo (angl. FIFO – first in first out) buferius ir nesudėtingus būsenų automatus buvo sukurta sistema jungianti neuroninį procesorių su RS232 moduliu, kuris savo ruoštu gali būti sujungtas su kompiuteriu ar kitų įrenginiu palaikančiu rs232 protokolą (19 pav.). Sintezės metu gautas NPW struktūrinis vaizdas pateiktas 3 priede.

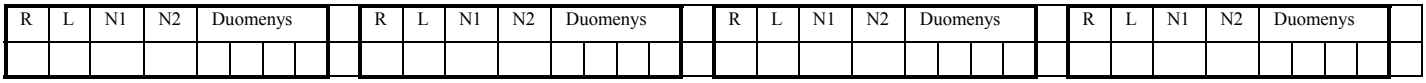
Duomenys siunčiami RS232 protokolu yra suskaldomi į keturias dalis, kiekviename pakete paliekant po 4 bitus neuroprocesoriaus valdymui ir likusius duomenų siuntimui .



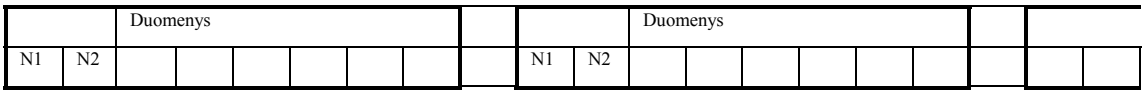
19 pav. Blokinė NPW schema

Naudojami du informacijos valdymo signalai - nustatymo į pradinę būseną R ir neuroprocesoriaus nustatymo į pradinę būsena signalą L. Paketo eilės numeriui žymėti skirti 2 bitai N1 ir N2.

Į RS232



Iš RS232



20 pav. Informacijos perdavimo paketų struktūra RS232 protokolu

Informacija iš neuroprocesoriaus skaldoma į 2 paketus po 8 bitus. Dėl to, kad neuroprocesoriaus išėjimo duomenų plotis yra 12 bitų, likę 4 bitai panaudojami paketo eilės numerio žymėjimui.

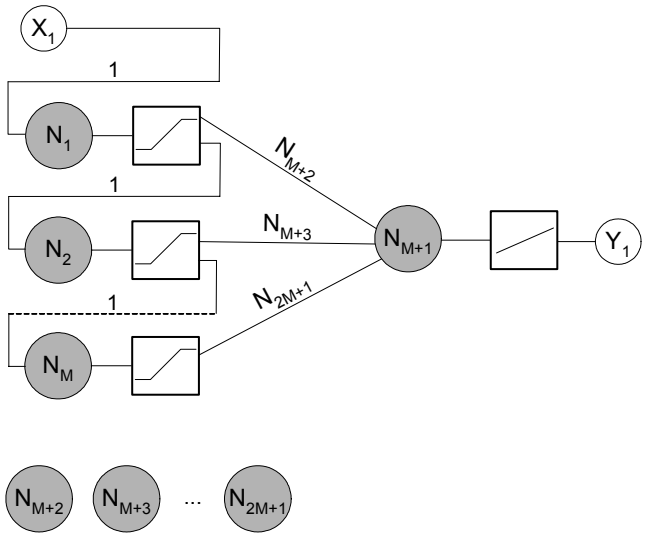
Informacijos perdavimui RS232 protokolas buvo pasirinktas dėl suderinamumo su personaliniu kompiuteriu. Nors įrenginio funkcionalumo testavimui greičio pakanka, šis protokolas yra lėtas ir netinkamas greitam neuroprocesoriaus darbui užtikrinti (3 priedas).

4. Eksperimentinė dalis

Eksperimento metu yra tiriamas neuroprocesoriaus veikimo teisingumas, analizuojami modeliuojamo neuroprocesoriaus skaičiavimo ciklai ir sintezuoto neuroprocesoriaus greičio ir lusto ploto parametrai. Analizuojant neuroninio procesoriaus veikimo teisingumą, yra lyginami matematinio neuroninio tinklo išėjimo duomenys su neuroprocesoriaus, realizuoto SystemC programavimo kalba, modeliavimo duomenimis ir sintezuoto neuroninio procesoriaus, pakrauto į FPGA įrenginį, išėjimo duomenimis.

4.1. Testiniai modeliavimo rezultatai

SystemC neuroprocesoriaus modelio veikimo teisingumui patikrinti naudojamas neuroninis tinklas panaudojantis pagrindines neuroprocesoriaus galimybes. Vienas iš standartinių neuroninio tinklo uždavinių yra adaptyvaus filtravimo taikymas. Tam, kad uždavinys butų paprastesnis, nėra realizuojamas grįžtamasis ryšys ir filtravimo parametrai nekeičiami.



21 pav. Neuroninio tinklo struktūra

Kaip parodyta paveiksle N_1-N_M neuronai atstoja įėjimų buferį, neuronas N_{M+1} yra išėjimo neuronas, kuris ir realizuoja filtro skaičiavimo algoritmą. Neuronai N_{M+2}-N_{2M+1} saugo neurono N_{M+1} svorius, arba kitaip filtravimo koeficientus. Pritaikius kitokias N_{M+2}-N_{2M+1} reikšmes arba juos keičiant neuroniniu tinklu, galima spręsti koreliacinius ar vidurkio skaičiavimo algoritmus. Tiesinė perdavimo funkcija N_{M+1} neuronui taikoma siekiant išgauti tiesinę priklausomybę visame skaičiavimo ruože.

Pavaizduoto 21 paveiksle neuroninio tinklo matematinė išraiška yra tokia:

$$Y_1 = F_2(\sum_{i=1}^{M} N_i \bullet N_{M+1+i}) \quad [12]$$

$$N_i=N_{i-1}\text{ , kai i=[2;M]} \quad [13]$$

$$N_1=X_1 \quad [14]$$

Neuroninio tinklo struktūra neuroprocesoriaus potinklyje atrodys taip:

4 lent. Neuroninio tinklo struktūra neuroprocesoriaus potinklio atminties masyve

	Atminties ląstelės tipas	**Atminties ląstelės adresas**	**Neuroninio tinklo reikšmė**	**Atminties ląstelių vėliavėlės**				**Atminties ląstelių reikšmės**
N_1	$A_{1,1}$	0	5	0	0	0	0	80
	$S_{1,1}$	1	32	1	0	0	0	520
	N_1	2	0	1	0	0	0	8
N2	$A_{2,1}$	3	8	0	0	0	0	128
	$S_{2,1}$	4	32	1	0	0	0	520
	N_2	5	0	1	0	0	0	8
N3	$A_{3,1}$	6	11	0	0	0	0	176
	$S_{3,1}$	7	32	1	0	0	0	520
	N_1	8	0	1	0	1	0	10
N4	$A_{4,1}$	9	14	0	0	0	0	224
	$S_{4,1}$	10	32	1	0	0	0	520
	N_4	11	0	1	0	0	1	9
N5	$A_{5,1}$	12	2	0	0	0	0	32
	$A_{5,1}$	13	22	0	0	0	0	352
	$A_{5,2}$	14	5	0	0	0	0	80
	$A_{5,2}$	15	23	0	0	0	0	368
	$A_{5,3}$	16	8	0	0	0	0	128
	$A_{5,3}$	17	24	0	0	0	0	384
	$A_{5,4}$	18	11	0	0	0	0	176
	$A_{5,4}$	19	25	0	0	0	0	400
	N_5	20	0	1	0	0	0	8
Išėjimas	O_5	21	1024	0	0	1	1	16386
Neurono svoriai	N_6	22	1024	1	0	0	0	16392
	N_7	23	-1024	1	0	0	0	49160
	N_8	24	1	1	0	1	0	26
	N_9	25	-1	1	0	0	0	65528

Matematinio neuroninio tinklo išėjimo reikšmės gautos naudojant 3 formulę. Kai D=32, o neuronų aktyvacijos slenkstis B=0.

5 lent. Neuroprocesoriaus testavimo rezultatai pakrovus neuroninį tinklą

X_1 Reikšmė	Matematinė neuroninio tinklo išėjimo reikšmė	Modeliuojamo neuroprocesoriaus išėjimo reikšmės	FPGA įrenginyje pakrauto neuroprocesoriaus išėjimo duomenys
1	0	0	0
2	1	1	1
3	-512	-512	-512
-1	1023	1023	1023
-2	-512	-512	-512
-3	1	1	1
2047	0	0	0
2046	0	0	0
2045	-512	-512	-512
-2047	1	1	1
-2046	0	0	0
-2045	1023	1023	1023
-3	0	0	0
-2	0	0	0
-1	512	512	512
0	0	0	0
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0

Iš 4 lentelės matyti, kad neuroprocesorius veikia lygiai taip kaip ir matematinis neuroninis tinklas. Taigi, galima daryti išvadą, kad neuroprocesoriaus, realizuoto FPGA technologijoje, funkcionalumas atitinka matematinį neuroninio tinklo modelį.

4.2. Sintezė

Modelis yra sintezuojamas FPGA architektūroje. Pasirinktas Sintezės įrankis yra Xilnix ISE programa, kuri naudoja XST Sintezatorių. Pasirinktas sintezės įrenginys Xilinx Virtex II XC2V40.

6 lent. Neuroprocesoriaus sintezės į Xilinx Virtex II XC2V40 pagrindiniai duomenys

Modulis	Alternatyvių vent. sk.	FPGA LUTs	FPGA Slices	Maksimalus dažnis, MHz
NPP	6093	193	107	100.644
npp_control	963	88	50	253.453
mem_1024x16	65539	0	0	81.920
NET	72572	273	152	81.920
rs_232	976	74	50	226.475
controler	132036	86	49	243.339
npw	205702	451	251	81.920

Remiantis sintezės rezultatais, matyti, kad maksimaliam neuroprocesoriaus dažniui daug įtakos turi atmintis mem_1024x16, nes jos dažnis yra mažiausias. Remiantis 17 paveiksle apskaičiuotu tiksiklio ciklų skaičiais vienai sinapsei galima apskaičiuoti maksimalų neuroprocesoriaus potinklio pajėgumą. Jei viena sinapsė apskaičiuojama per 6 tiksiklio ciklus, tai apskaičiuojamų sinapsių skaičius per sekundę esant 80 MHz dažniui yra
13.3 M sinapsių/s.

5. Išvados

- Neuroprocesoriai, galintys apdoroti grįžtamojo ryšio neuroninius tinklus, yra pritaikomi plataus spektro uždaviniams spręsti;
- pasiūlytas apibendrintas neuroprocesoriaus modelis, apibūdina literatūros šaltiniuose aprašytų neuroprocesorių daugumą;
- sudarytą matematinį modelį galima taikyti tiesioginio sklidimo ir grįžtamojo ryšio neuroniniams tinklams;
- apibendrinus neurokomponentus, neuroprocesorius yra pritaikomas įvairaus dydžio FPGA įrenginiams;
- sukurtas neuroprocesoriaus prototipas turi 12 bitų duomenų plotį;
- neuroprocesoriuje naudojant tiesinę ir tiesinę ribojančią perdavimo funkciją, sutaupomas integrinės schemos plotas.
- vienai sinapsei apskaičiuoti neuroprocesoriuje reikia nuo 6 iki 12 tiksiklio ciklų;
- neuroprocesoriaus prototipo modeliavimo rezultatai parodė, kad sukurto neuroprocesoriaus funkcionalumas atitinka matematinį neuroninio tinklo modelį.
- sintezės rezultatai parodė, kad maksimalus neuroprocesoriaus dažnis Xilinx Virtex II XC2V40 įrenginyje yra 82 MHz , tai laidžia pasiekti iki 13M sinapsių per sekundę skaičiavimo pajėgumą.

6. Literatūra

[1] Antsaklis P., M. Passino K. Introduction to Intelligent Control systems with high degree of autonomy. OSU Collaborative Center of Control Science [interaktyvus]. [žiūrėta 2004-12-12]. Prieiga per internetą: http://www.ece.osu.edu/~passino/index.html

[2] Liao Y. Neural Networks in Hardware A Survey: California: Department of Computer Science, University of California [interaktyvus]. 2001 [žiūrėta 2004-12-12]. Prieiga per internetą: http://ailab.das.ucdavis.edu/~yihua/research/.

[3] Anderson D., McNeil G. Artificial Neural Networks Technology. Iš DACS [interaktyvus]. 1992, rugpjūtis [žiūrėta 2005-04-03].
Prieiga per internetą: http://www.dacs.dtic.mil/techs/neural/neural_ToC.

[4] Asaro P. Neural Networks. CITES NetFiles [interaktyvus]. [žiūrėta 2005-04-03]. Prieiga per internetą: https://netfiles.uiuc.edu/asaro/www/writing/neural_networks.html.

[5] Lindsey C. Neural. Networks in Hardware.Lecture A: Overvie. Sweden: Royal Institute of Technology Stockholm [interaktyvus], [žiūrėta 2005-04-15]. Prieiga per internetą: http://www.particle.kth.se/~lindsey.

[6] Kėvelaitis E., Illert M., Hultborn H. Žmogaus fiziologija. Abraitis ir kt. 1999. 478 p.

[7] Hochreiter, Sepp, Schmidhuber, Juergen, Long Short-Term Memory, Neural Computation. Iš IDSIA. [interaktyvus]. 1997 [žiūrėta 2005-5-1]. Prieiga per internetą: http://www.idsia.ch/NNcourse/lstm.html

[8] Schmidhuber J. Recurrent Neural Networks. Iš IDSIA. [interaktyvus]. 2004. [žiūrėta 2005-5-1]. Prieiga per internetą: http://www.idsia.ch/~juergen/rnn.html

[9]A236™ Parallel Video DSP Chip from oxford Micro Devices, Data sheet Summary. Iš Oxford Micro Devices [interaktyvus]. [žiūrėta 2004-12-12]. Prieiga per internetą: http://www.oxfordmicrodevices.com/.

[10] Jones S., Sammut K. Learning in Systolic Neural Network Engines Department of electronic and electrical engineering. Loughborough University of Technology. UK. [interaktyvus]. [žiūrėta 2005-5-1]. Prieiga per internetą: http://www.bath.ac.uk/engineering/group/publications/cd_papers/

[11] HAGEN - The Heidelberg AnaloG Evolvable Neural Network. Iš University of Heidelberg-Kirchhnoff Institute for Physics [interaktyvus]. [žiūrėta 2004-12-12] Prieiga per internetą: http://www.uni-heidelberg.de/index_e.html

[12] Clarke. P. Startup implements silicon neural net in Learning Processor. Iš EE Times [interaktyvus]. 1999, vasaris[žiurėta 2005-02-03]. Prieiga per internetą: http://www.eet.com/story/OEG19990914S0033].

[13] Aryashev S., Bobkov S., Sidorov E., Yudin I. Parallel FPGA Processor Card for Distributed Information Processing. Institute for System Studies, Russian Academy of Sciences [interaktyvus].
[žiūrėta 2004-12-12]. Prieiga per internetą: http://www.omimo.be/magazine/99q2/Russian.pdf

[14] Konfigūruojamoji kompiuterinė įranga. UAB "Ryšių technikos naujienos". M. K. Čiurlionio 7/1, 03104 Vilnius. Žurnalas "Ryšių technikos naujienos" [interaktyvus]. [žiūrėta 2004-12-12]. Prieiga per internetą: http://www.rtn.lt/rtn/9904/konfig.html

[15] UAB "Ryšių technikos naujienos". M. K. Čiurlionio 7/1, 03104 Vilnius. Žurnalas "Ryšių technikos naujienos" . 2003 m. Nr. 2. Šalin mikroprocesorius? [interaktyvus]. [žiūrėta 2004-12-12]. Prieiga per internetą: http://www.rtn.lt/rtn/0302/mikroprocesorius.html

[16] Zhu J., Sutton P. FPGA Implementations of Neural Networks – a Survey of a Decade of Progress School of Information Technology and Electrical Engineering. The University of Queensland, Brisbane, Queensland 4072, Australia. [interaktyvus]. [žiūrėta 2004-12-12]. Prieiga per internetą: http://www.itee.uq.edu.au/~peters/papers/

[17] Edwards S. Synopsys, Inc. Mountain View, California. Design Languages for Embedded Systems. Columbia University, Department of Computer Science [interaktyvus].
[žiūrėta 2004-12-12]. Prieiga per internetą: http://www1.cs.columbia.edu/~sedwards

[18] Panda P. SystemC A modeling platform supporting multiple design abstractions. Synopsys Inc.700 E. Middlefield Rd. Mountain View, CA 94043, USA [interaktyvus].
[žiūrėta 2004-12-12]. Prieiga per internetą:
http://www.cse.iitd.ernet.in/~panda/SYSTEMC/Tutorials

[19] Edwards S. The Challenges of Hardware Synthesis from C-like Languages. Department of Computer Science, Columbia University, New York. [interaktyvus]. [žiūrėta 2004-12-12]. Prieiga per internetą: http://www1.cs.columbia.edu/~sedwards/papers/

[20] C in Hardware and SoC Design. Iš Celoxica inc. [interaktyvus]. [žiūrėta 2005-5-1]. Prieiga per internetą: http://www.celoxica.com/technology/c_design/hardware.asp.

[21] Schoenauer T., Jahnke A., Roth U., Klar H. Digital Neurohardware.Principles and Perspectives. Institute of Microelectronics, Technical University of Berlin [interaktyvus].
[žiūrėta 2004-12-12]. Prieiga per internetą: http://mikro.ee.tu-berlin.de/~spinnsoc/.

[22] Eppler W., Fischer T., Gemmeke H., Menchikov A. High Speed Neural Network Chip for Trigger Purposes in High Energy Physics. Iš IEEE Xplore [interaktyvus]. [žiūrėta 2004-12-12]. Prieiga per internetą: http://ieeexplore.ieee.org.

[23] Fieres J., Philipp S., Meier K., ir kt. A Platform for Parallel Operation of VLSI Neural Networks. University of Heidelberg, Kirchhoff Institute for Physics [interaktyvus]. [žiūrėta 2004-12-12]. Prieiga per internetą: http://www.kip.uni-heidelberg.de/vision

7. Terminų ir santrumpų sąrašas

FPGA – programuojama ventilių matrica arba programuojamųjų sklendžių matrica (angl. *field-programmable gate array*)

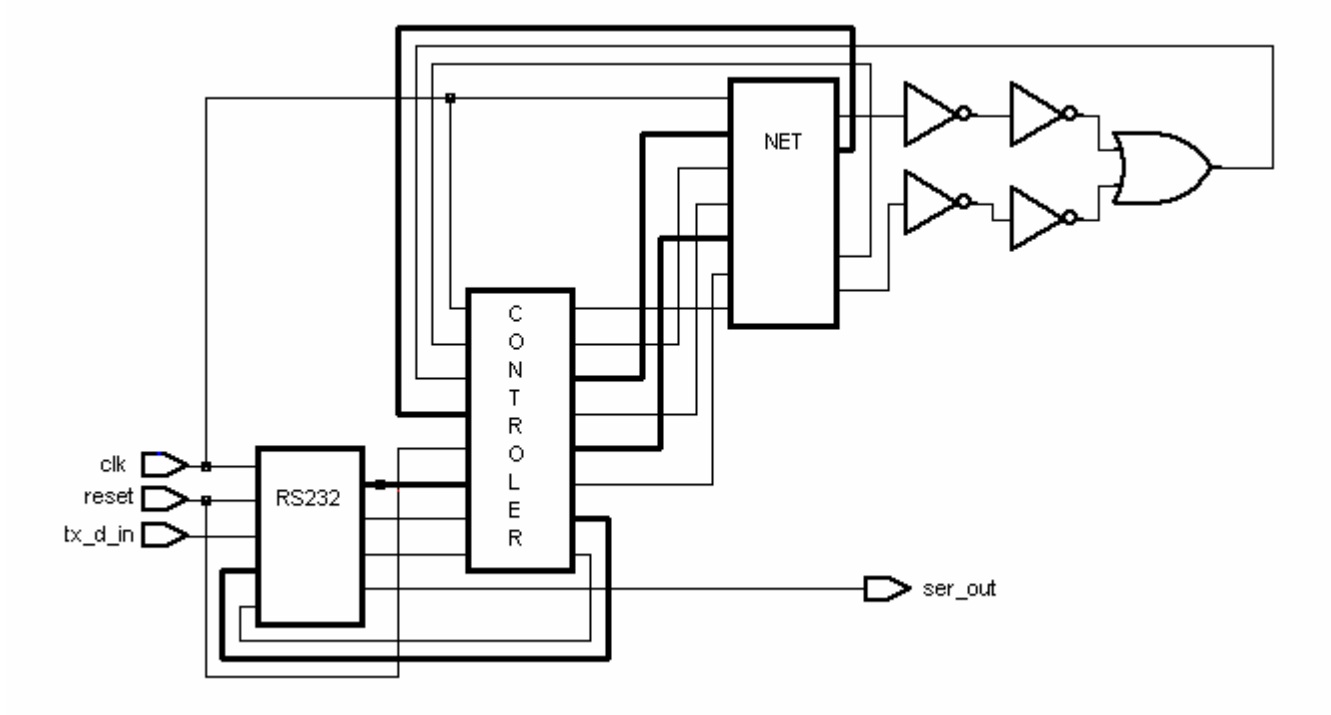
Fuzzy – neaiškus, neraiškus, miglotas

DNT– dirbtiniai neuroniniai tinklai

VHDL – aukšto lygmens aparatūros aprašymo kalba (angl. *Very High Hardwar Description Language*)

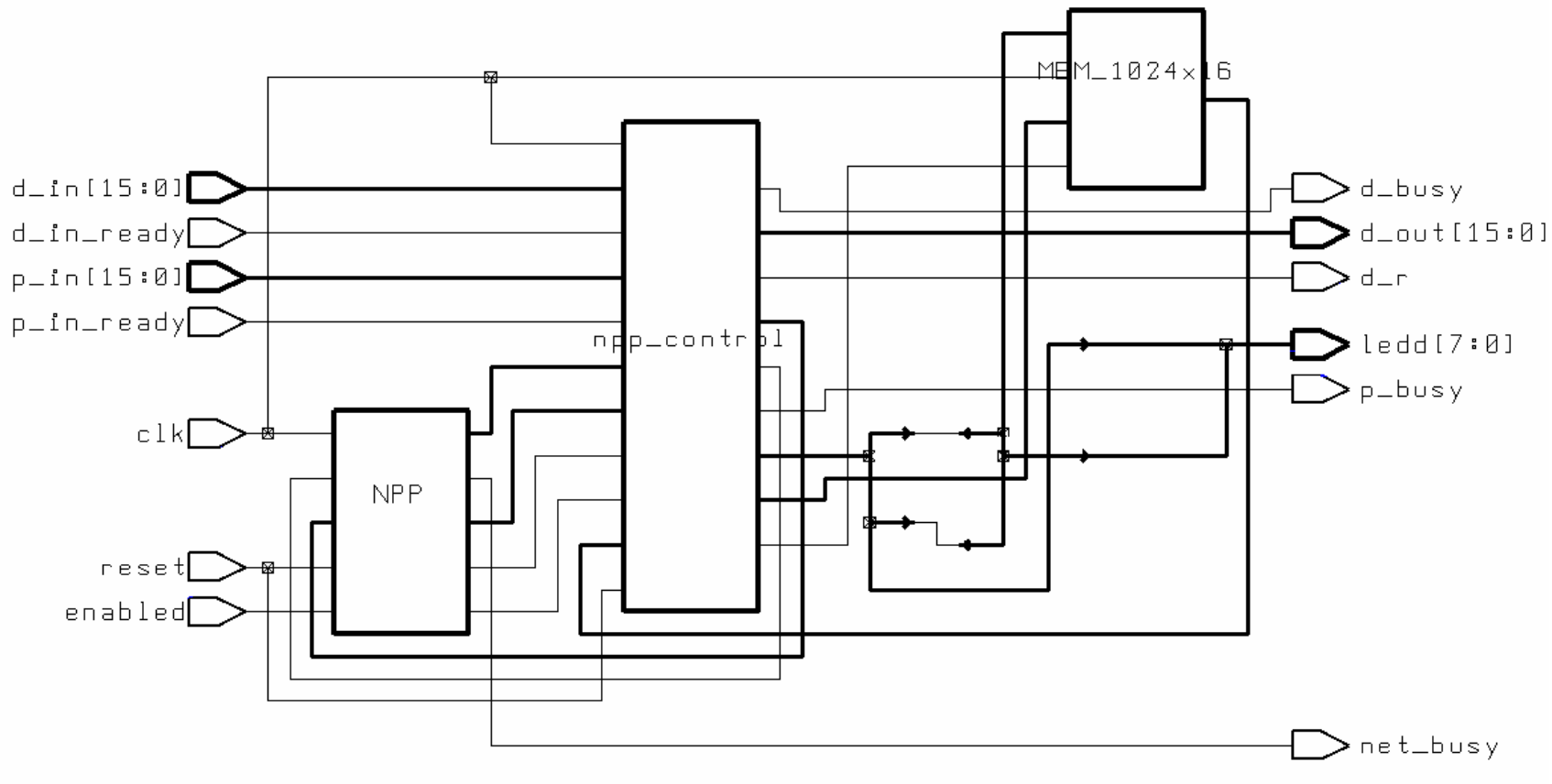
8. Priedai

8.1. 1 priedas



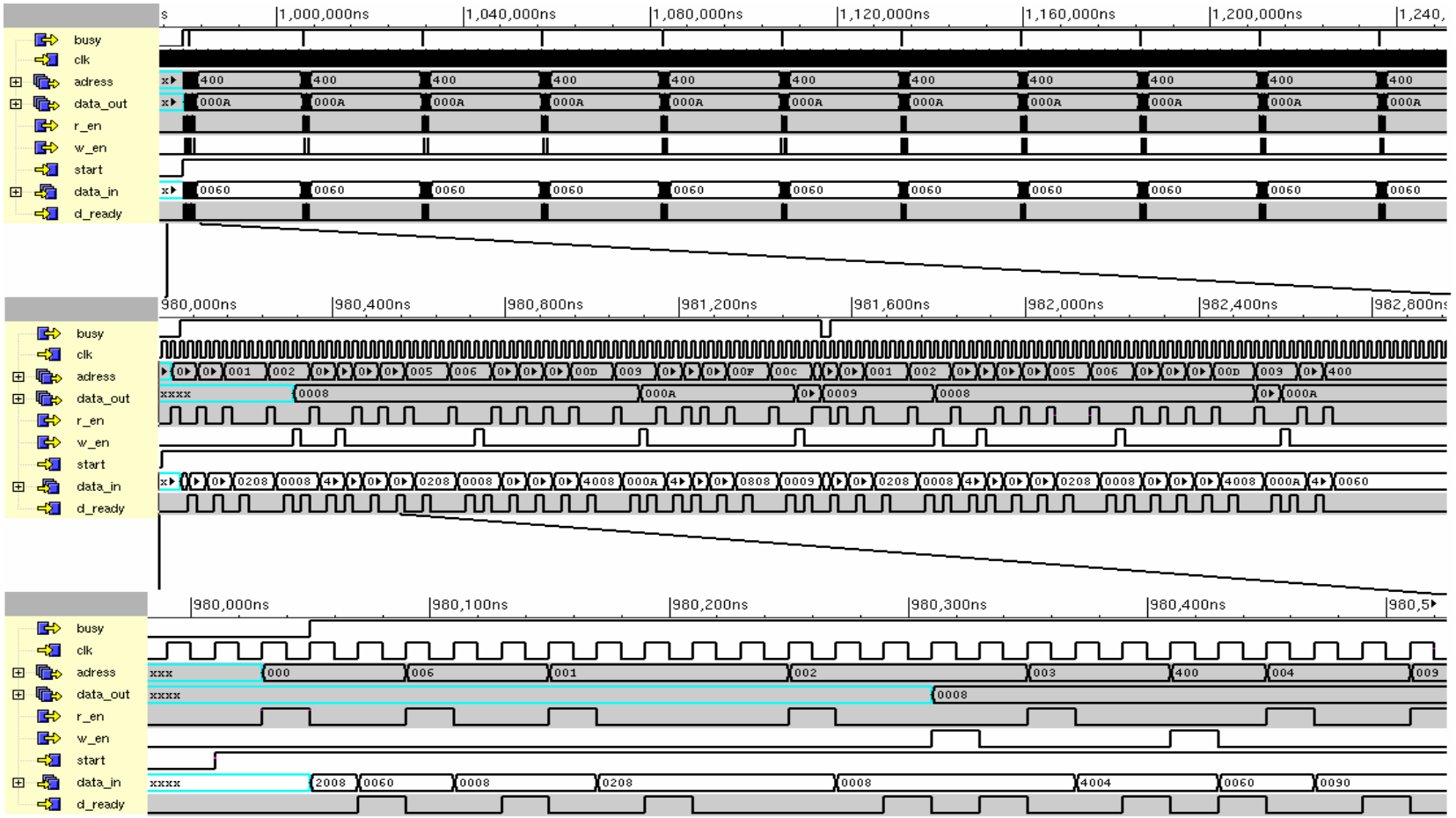
22 pav. Sintezuoto NPW struktūrinis vaizdas

8.2. 2 priedas



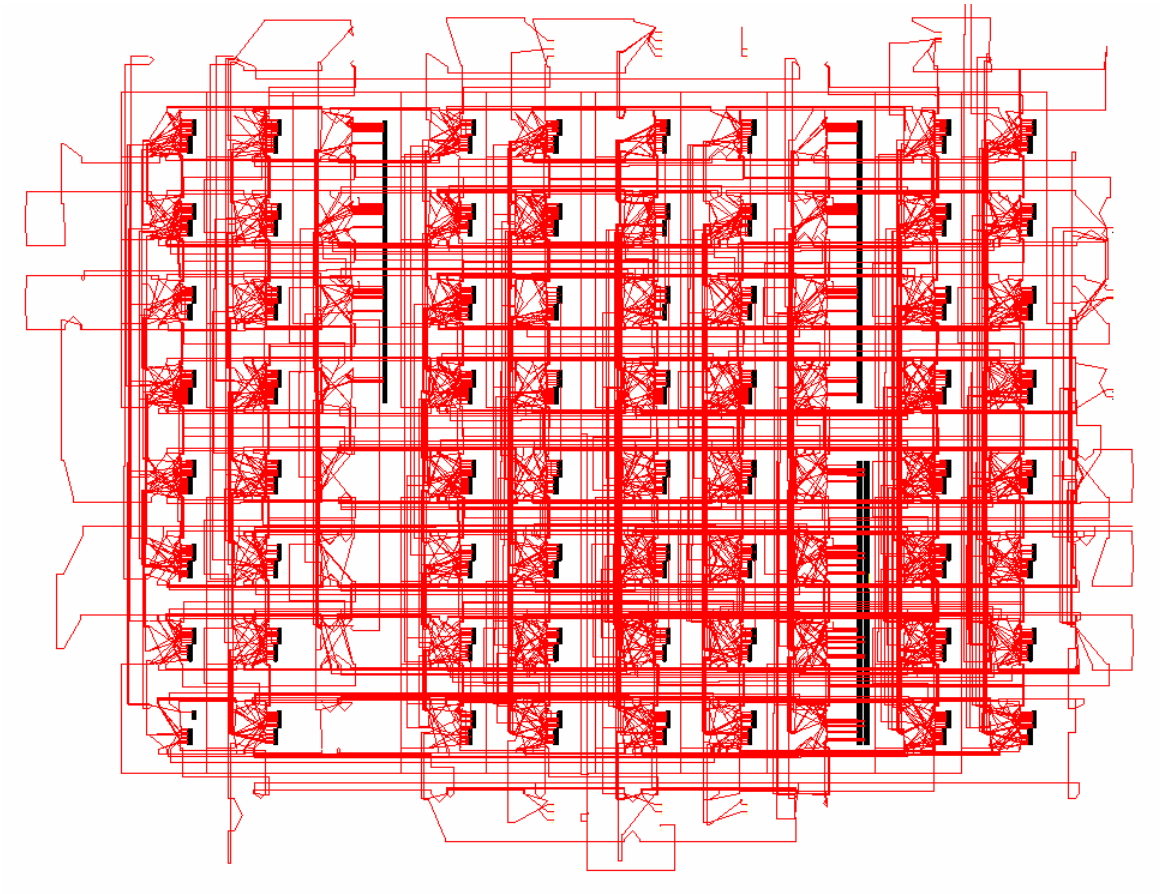
23 pav. Sintezuoto NPP struktūrinis vaizdas

8.3. 3 priedas



24 pav. Laikinės NPP bloko diagramos

8.4. 4 priedas



25 pav. Sutrasuoto FPGA vaizdas užkrovus neuroprocesorių

8.5. Kiti priedai

Priedai pateikiami kompaktiniame diske:

- SystemC išeities tekstas
- XILINX ISE PROJEKTAS
- Šio darbo skaitmeninė kopija