VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV MIKROELEKTRONIKY
DEPARTMENT OF MICROELECTRONICS

IMPLEMENTACE PCS A PMA PODVRSTVY 50 GB/S ETHERNETU V FPGA
FPGA IMPLEMENTATION OF PCS AND PMA SUBLAYER OF 50GB/S ETHERNET

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Michal Suchanek**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Marek Bohrn, Ph.D.**
SUPERVISOR

BRNO 2019

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní obor **Mikroelektronika a technologie**
Ústav mikroelektroniky

Student: Michal Suchanek ***ID:*** 186455
Ročník: 3 ***Akademický rok:*** 2018/19

NÁZEV TÉMATU:

Implementace PCS a PMA podvrstvy 50 Gb/s Ethernetu v FPGA

POKYNY PRO VYPRACOVÁNÍ:

Seznamte se s architekturou FPGA na zvolené akcelerační kartě a se standardem 25/50 Gigabit Ethernet Consortium definujícím 50 Gb/s Ethernet. Podrobně prostudujte požadavky na PCS a PMA podvrstvy fyzické vrstvy Ethernetu v režimu 50GBASE-R. Navrhněte jednotky implementující navržené PCS a PMA podvrstvy v FPGA na zvolené akcelerační kartě.
Implementujte navrženou jednotku a ověřte funkčnost implementované jednotky na zvolené akcelerační kartě. V závěru diskutujte dosažené výsledky a zamyslete se nad možnostmi optimalizace spotřeby zdrojů na FPGA.

DOPORUČENÁ LITERATURA:

Podle pokynů vedoucího práce

Termín zadání: 4.2.2019 ***Termín odevzdání:*** 30.5.2019

Vedoucí práce: Ing. Marek Bohrn, Ph.D.
Konzultant: Ing. Štěpán Friedl, CESNET, z.s.p.o.

doc. Ing. Jiří Háze, Ph.D.
předseda oborové rady

UPOZORNĚNÍ:
Autor bakalářské práce nesmí při vytváření bakalářské práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

Abstrakt

Cílem této bakalářské práce je seznámit se se standardem 25/50 Gigabit Ethernet Consortium, jenž definuje 50Gb/s Ethernet. Prostudovat specifikace pro PCS a PMA podvrstvy fyzické vrstvy Ethernetu v režimu 50GBASE-R. Podle těchto specifikací navrhnout a implementovat zmiňované podvrstvy PCS a PMA v jazyce VHDL pro obvody FPGA na akcelerační kartu. Ověřit funkčnost fyzické vrstvy na zvolené akcelerační kartě.

Klíčová slova

FPGA, ethernet, VHDL, fyzická vrstva

Abstract

The main goal of this thesis is to familiarize with 25/50 Gigabit Ethernet Consortium standard, which defines 50Gb/s Ethernet. Study about PCS and PMA sublayer specifications for Ethernet physical layer in 50GBASE-R mode. Describe and implement mentioned PCS and PMA sublayers in VHDL language for FPGA circuits and selected acceleration card. Verify correct functionality of physical layer through tests on given acceleration card.

Keywords

FPGA, ethernet, VHDL, physical layer

Bibliografická citace:

SUCHANEK, Michal. *Implementace PCS a PMA podvrstvy 50 Gb/s Ethernetu v FPGA*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2019. 51 s.. Bakalářská práce. Vedoucí práce Ing. Marek Bohrn, Ph.D..

Prohlášení

„Prohlašuji, že svou bakalářskou práci na téma Implementace PCS a PMA podvrstvy 50 Gb/s Ethernetu v FPGA jsem vypracoval samostatně pod vedením vedoucí/ho bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené semestrální práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne: …… ……...
podpis autora

Poděkování

Děkuji vedoucímu bakalářské práce Ing. Marku Bohrnovi, Ph.D. a konzultantovi bakalářské práce Ing. Štěpánu Friedlovi za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé bakalářské práce.

V Brně dne: …… ……...
podpis autora

Obsah

Seznam obrázků **7**

Seznam tabulek **8**

Úvod **9**

1. Obvody FPGA **10**

1.1 Architektura FPGA **10**

Vstupně-výstupní bloky *10*

Logické buňky *10*

Propojovací matice *13*

Speciální funkční bloky *14*

2. Ethernet **15**

2.1 Fyzická vrstva **15**

2.2 PCS podvrstva **16**

MII sběrnice *16*

2.3 FEC podvrstva **19**

2.4 PMA a PMD podvsrtvy **19**

3. Hardwarové specifikace implementace **20**

3.1 Multi-Gigabitové Transcievery **20**

3.2 FP transcievery **20**

QSFP28 ..... *21*

CFP4 ..... *21*

3.3 Cílové síťové karty **21**

4. Návrh fyzické vrstvy **23**

4.1 Parametry návrhu **23**

4.2 PCS podvrstva **23**

4.3 PMA a PMD podvrstva **24**

4.4 Řídicí blok **25**

MI32 sběrnice *26*

5. Komponenty **27**

5.1 Top modul **27**

Porty rozhraní fyzické vrstvy *29*

Architektura top modulu *30*

5.2 Management (Řídicí blok) 31
5.3 PCS vrstva 31
5.4 PMA Vrstva pro UltraScale+ 33
Generování GTY transcieveru 33
Zpracování dat a bitový multiplex 34
Implementace block lock logiky 35
Řízení hodinových signálů 36
5.5 PMA vrstva pro Virtex 7 37
Generování GTZ transcieveru 38
Řízení hodinových signálů 40
Zpracování dat a bitový multiplex 40
6. Syntéza a testování 42
6.1 Syntéza 42
Využití zdrojů 42
Časová analýza 43
6.2 Testování a ověření funkčnosti 43
Simulace a verifikace obvodu 44
Testování přes Spirent TestCenter 45
7. Závěr 49
Literatura 50
Seznam symbolů, veličin a zkratek 51

SEZNAM OBRÁZKŮ

Obr. 1.1: Architektura FPGA obvodu [6] 11
Obr. 1.2: Konfigurovatelný logický blok [6] 12
Obr. 1.3: Zjednodušená struktura logického řezu [6] 13
Obr. 2.1: Blokové schéma fyzické vrstvy 15
Obr. 2.2: Diagram matematické funkce scrambleru [1] 17
Obr. 2.3: Rozdíl režimů pro fyzickou vrstvu [3] 18
Obr. 3.1: FPGA karty pro implementaci 50G ethernetu [4] 22
Obr. 4.1: Funkce multiplexeru 4:2 a 2:4 25
Obr. 4.2: Příklad zápisové a čtecí transakce MI32 sběrnice 26
Obr. 5.1: Blokové schéma top modulu fyzické vrstvy 30
Obr. 5.2: Proces rozdělení výstupních dat GTY transcieveru na 66 bitové bloky 35
Obr. 5.3: Zvolené kanály GTZ transcieveru včetně zdrojů hodinových signálů 39
Obr. 5.4: Systém výstupních/vstupních dat GTZ transcieveru na rozhraní PCS a PMA vrstvy 41
Obr. 6.1: Propojení portů Spirent modulu a akcelerační karty 44
Obr. 6.2: Výstup z konzole serveru s připojenou akcelerační kartou 46
Obr. 6.3: Závislost odeslaných a přijatých ethernetových rámců za jednotku času na délce rámce 48

SEZNAM TABULEK

Tab. 3.1: Porovnání specifikací FPGA karet od firmy Netcope 22
Tab. 5.1: Seznam použitých a vytvořených komponent 32
Tab. 6.1: Využití zdrojů fyzické vrstvy pro jednoportovou verzi UltraScale+ 42
Tab. 6.2: Výsledky časové analýzy implementace síťové karty 43
Tab. 6.3: Výsledky testů fyzické vrstvy pro různé délky rámců (Spirent TestCenter) 47
Tab. 6.4: Výsledky testů fyzické vrstvy pro různé délky rámců (výstup z konzole) 47

Úvod

Podle modelu ISO/OSI pracuje internet jako systém na několika vrstvách, kde každá vrstva má svou specifickou funkci. Nejnižší z těchto vrstev je vrstva fyzická, která se stará o přenos a zpracování bitů dát na nejnižší možné úrovni. Vzhledem k tomu, že trend rychlostí internetu jde stále kupředu, je třeba vyvíjet a připravovat komponenty, které umožňují zvládat čím dál tím vyšší rychlosti ethernetu. Avšak je nutné myslet také na to, že existuje několik režimů ve kterém může ethernet a fyzická vrstva pracovat.

Předmětem této práce je návrh a implementace fyzické vrstvy 50G etherentu v jazyce VHDL pro síťové karty s čipy FPGA pro dvě různé architektury těchto obvodů, a to konkrétně architektury UltraScale+ a Virtex 7. Návrh vychází ze standardu 25/50 Gigabit Ethernet Consortium jenž definuje 50 Gb/s ethernet. Vytvoření popisu je zadáním projektu pro firmu CESNET, z. s. p. o.

Součástí práce je rozbor a podrobný popis problematiky fyzické vrstvy 50G ethernetu, jeho podvrstev PCS a PMA, a funkcí jež tyto vrstvy realizují. Dále se práce věnuje samotnému návrhu fyzické vrstvy a jeho podvrstev v jazyce VHDL pro obvody FPGA. V další části práce jsou pak detailně popsány použité komponenty včetně jejich realizace v jazyce VHDL. Součástí práce je také postup generování gigabitových transcieverů pro obě architektury obvodů FPGA. V závěru práce jsou shrnuty výsledky fyzického testování na akcelerační kartě.

1. OBVODY FPGA

Následující kapitola je věnována programovatelným logickým obvodům FPGA (Field Programmable Gate Array). Obvody FPGA se řadí do skupiny obvodů s označením PLD, neboli Programmable Logic Device. PLD jsou programovatalné logické obvody jejichž funkci a aplikaci lze měnit (programovat). Touto vlastností se velmi podobá například mikrokontrolerům, kdy se přeprogramováním softwaru dá změnit výsledná funkčnost kontroleru. V případě PLD obvodů je však přeprogramována struktura hardwaru, přesněji jejich propojení a využití jednotlivých prvků. K tomuto účelu slouží nízkoúrovňové programovací jazyky jako VHDL (VHSIC Hardware Description Language), Verilog a další. Jazyk ve kterém bude implementována fyzická vrstva ethernetu bude VHDL. [7]

1.1 Architektura FPGA

Architektura FPGA znázorněna na Obr. 1.1, je tvořena několika základnímy prvky:

- konfigurovatelné vstupně-výstupní bloky
- konfigurovatelné logické buňky
- programovatelná propojovací matice
- speciální funkční bloky (blokové paměti, sériové transcievery, DSP bloky, ...)

Vstupně-výstupní bloky

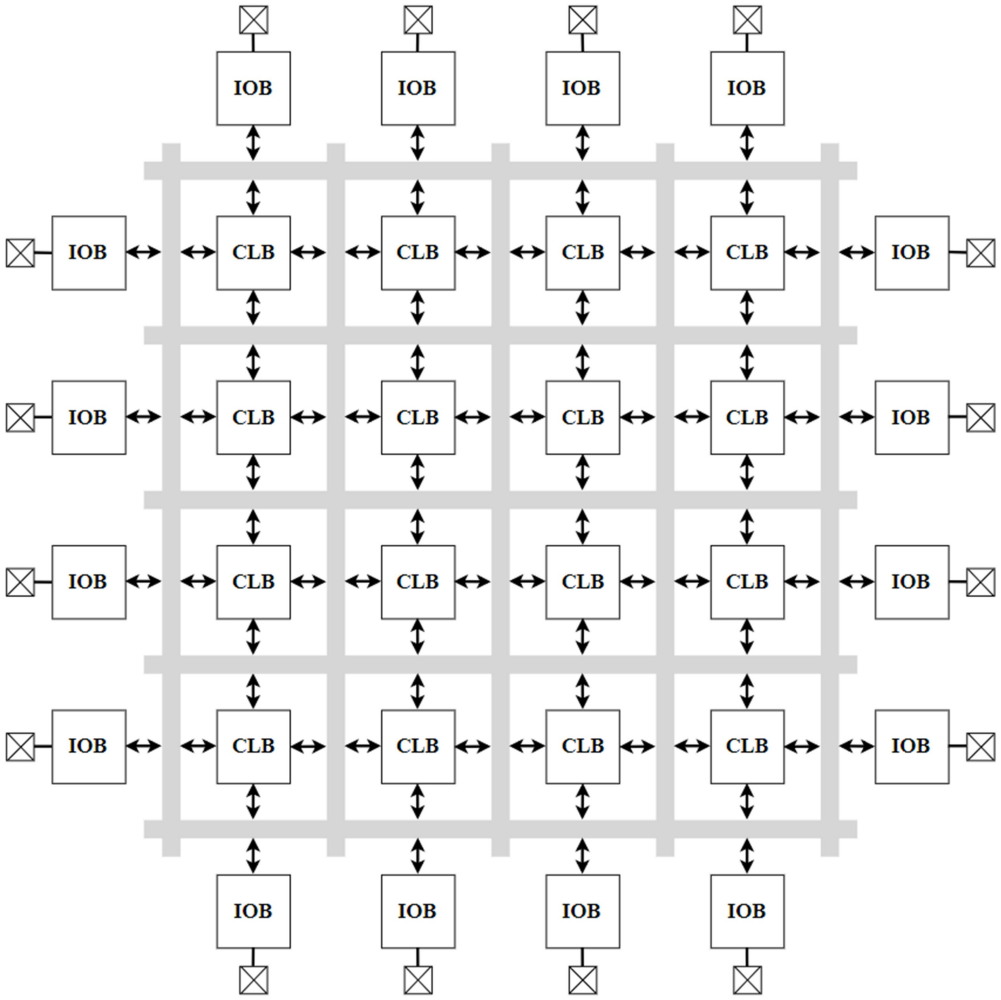
Vstupně-výstupní bloky, nebo také IOB (Input-Output Block), jsou základním prvkem na rozhraní FPGA přes které prochází každý signál vstupující nebo vystupující z FPGA. IOB obsahují vstupní a výstupní registry, zpožďovací linky, budiče a příjmače, obvody impedančního přizpůsobení a ochranné obvody.

Logické buňky

Nejzákladnějším prvkem FPGA obvodů jsou konfigurovatelné logické buňky, také označovány jako CLB (Configurable Logic Block). Slouží k realizaci kombinačních a sekvenčních logických funkcí. Základní struktura logického bloku je znázorněna na Obr. 1.2. Logické buňky obsahují zpravidla dva logické řezy, z anglického "Slice".

Logický řez (Slice) je ve zjednodušené podobě znázorněn na Obr. 1.3 a je složen z několika základních částí:

- náhledových tabulek (LUT)
- registrů
- multiplexerů
- řeťězců šíření přenosu

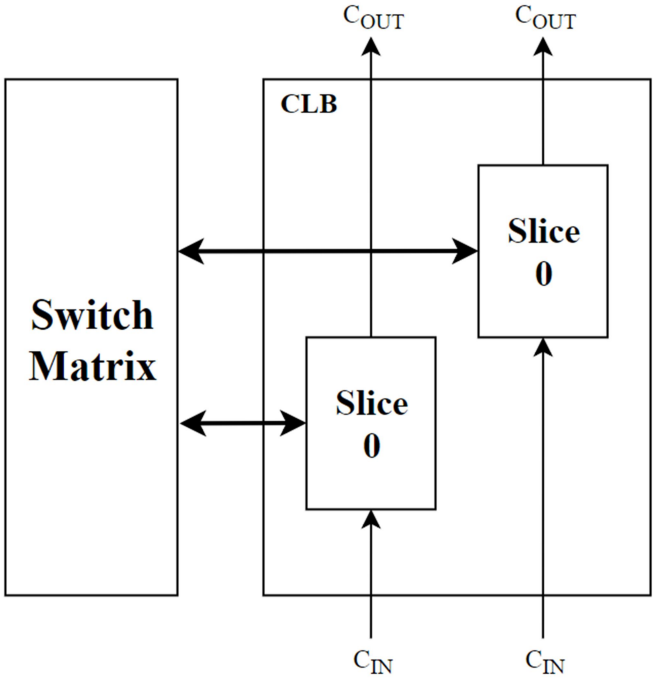


Obr. 1.1: Architektura FPGA obvodu [6]

Náhledové tabulky (LUT – Look-Up Table) jsou prioritně používány pro realizaci kombinační logiky. V zásadě jde o paměť RAM jenž je vždy konfigurována pro danou logickou funkci. Zpravidla se jedná o 4-vstupové jednotky s jedním výstupem. Novější architektury FPGA obashují až 6ti vstupové LUT.

Pro sekvenční část logických bloku jsou zde k dispozici registry. Jedná se o klopný obvod typu D. Konfigurací lze zvolit, zda budou registry využívat signály jako clock-enable (CE), set/reset (S/R), polaritu a další.

Další důležitou součástí logických řezů jsou programovatelné multiplexery. Na Obr. 1.3 je znázorněn 3-vstupový multiplexor. Pro vícevstupové multiplexory se logika realizuje kombinací náhledových tabulek v jednom logickém řezu.



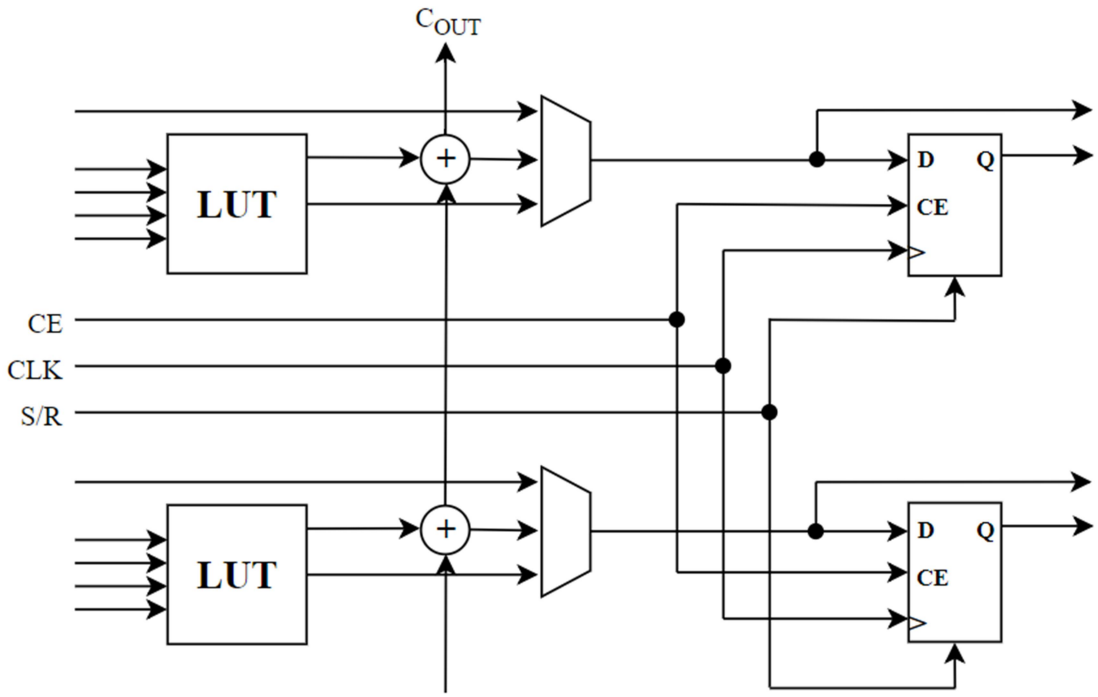
Obr. 1.2: Konfigurovatelný logický blok [6]

V neposlední řadě obsahují logické řezy řetězce **šíření přenosu** (CIN, COUT), nebo také z anglického "carry chain". Základní funkionalitou je vytvoření aritmetických funkcí (např. sčítačky). Podstatou těchto prvků je přímé propojení mezi logickými bloky v daném sloupci, oproti ostatním signálům, které jsou propojeny přes globální propojovací matici. Výhodou tohoto propojení je mnohem větší rychlost přenosu.

Propojovací matice

K propojení jednotlivých logických bloků FPGA slouží 3 typy propojovacích vodičů:

- **single-length** – přímé propojení sousedních logických buňek, zaručeno nejmenší zpoždění
- **double-length** – do středních vzdáleností se využívají spojovací matice PSM (Programmable Switch Matrix).
- **longlines** – horizontální a vertkální vodiče vedeny napříč celou součástkou pro kritické signály na dlouhou vzdálenost.



Obr. 1.3: Zjednodušená struktura logického řezu [6]

Speciální funkční bloky

Mezi speciální funkční bloky se řadí blokové paměti, sériové transcievery, DSP bloky a další...

Blokové paměti jsou zpravidla dvouportové statické paměti s kapacitou až desítek kB. Jsou vysoce univerzální a lze je využít jako jednoportová nebo dvouportová paměť typu RAM nebo ROM. Lze nastavit také šířku datové sběrnice (1, 2, 4, 8, 16, 32 nebo 64).

Sériové transcievery slouží pro realizaci velmi rychlých sériových rozhraní. Transcievery v sobě obsahují extraktor hodinového signálu, diferenciální budič a příjmač, serializer a deserializer, kodér a dekodér a další pomocné obvody.

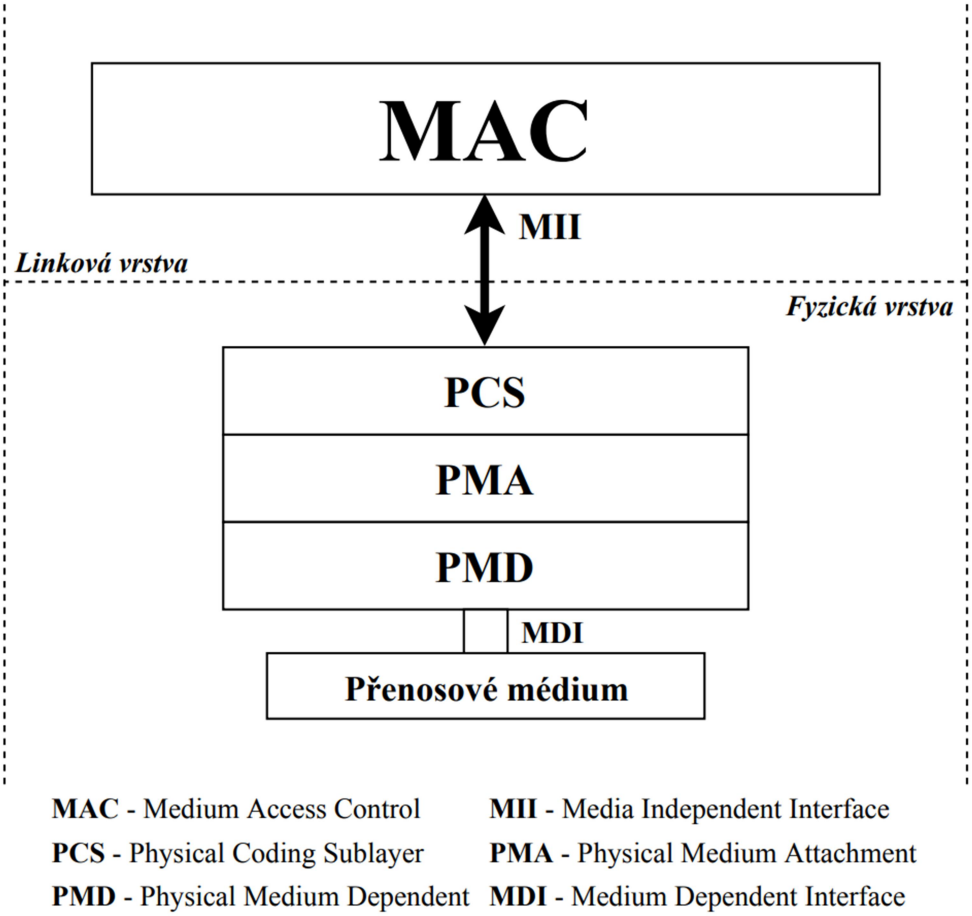
DSP bloky umožňují realizaci složitějších matematických operací. Hlavní výhodou a účelem DSP bloků je, že danou matematickou operaci dokáže zvládnout v jednom hodinovém cyklu.

2. ETHERNET

Ethernet je síťový protokol pracující na 1. a 2. vrstvě modelu ISO/OSI, tedy fyzické a linkové. Normy IEEE 802 obsahují standardy pro fyzickou a linkovou vrstvu. Konkrétně norma IEEE 802.3 specifikuje standard ethernetu pro lokální síť využívající společné komunikační médium. [2]

2.1 Fyzická vrstva

Hlavním úkolem fyzické vrstvy je zajistit přenos bitů mezi odesílatelem a příjemcem. Z linkové vrstvy dostává fyzická vrstva datový rámec, který musí zakódovat a převést do bitové přenositelné podoby. Blokové schéma fyzické vrstvy je znázorněno na Obr. 2.1.



Obr. 2.1: Blokové schéma fyzické vrstvy

Znázorněné blokové schéma a následující informace o fyzických podvrstvách vycházejí ze standardu pro 40G ethernet od IEEE 802.3 a pro 50G ethernet od Consortium.

Mezi podvrstvy fyzické vrstvy patří:

- PCS – Physical Coding Sublayer
- FEC – Forward Error Correction
- PMA – Physical Medium Attachment
- PMD – Physical Medium Dependent

2.2 PCS podvrstva

Úkolem PCS podvrstvy je zakódovat data z linkové vrstvy do 66 bitových bloků (64B/66B kódování), vložit zarovnávací značky (aligment markers) a odeslat alší podvrstvě. Toto platí pro vysílací cestu. V případě příjmací cesty je postup opačný, tedy najít a odstranit zarovnávací značky, dekódovat do podoby MII (Media Independent Interface) sběrnice a odeslat do linkové vrstvy.

MII sběrnice

Komunikační sběrnice mezi linkovou vrstvou a fyzickou vrstvou (PCS podvrstvou). MII sběrnice zpravidla obsahují datový, řídicí a hodinový signál. Značení MII sběrnic se třídí podle rychlosti přenosu a šířky sběrnice, příklady značení je následující:

- **GMII** – Gigabit Media Independent Interface
- **XGMII** – 10 (římsky "X") Gigabit Media Independent Interface
- **XLGMII** – 40 (římsky "XL") Gigabit Media Independent Interface
- **CGMII** – 100 (římsky "CG") Gigabit Media Independent Interface
- **CDGMII** – 400 (římsky "CD") Gigabit Media Independent Interface

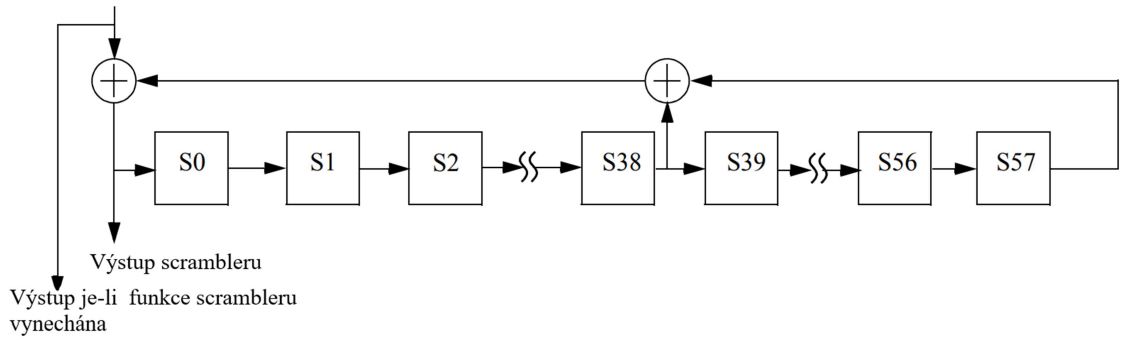
Datový signál je označován jako TXD (Transmit – vysílací) nebo RXD (Recieve – příjmací). Přenáší hlavní data rozdělena to 8-bitových bloků, kdy každá osmice je označena jako linka. Sekvence ve kterém je vysílán rámec je následující:

<inter-frame><preamble><sfd><data><efd>

- **inter-frame** – mezirámec obsahující "Idle"znaky při nečinnosti sběrnice (hexadecimálně "07")
- **preamble** – preambule vysílaná před začátkem rámce. Řídicí znak zarovnán na první linku sběrnice (první osmice bitů), následuje 6 stejných osmic preambule (binárně "10101010")
- **sfd** – osmice bitů následující hned po preambuli (hexadecimálně "FB")
- **data** – obsahují ethernetový rámec
- **efd** – reprezentuje ukončovací řídicí znak (hexadecimálně "FD")

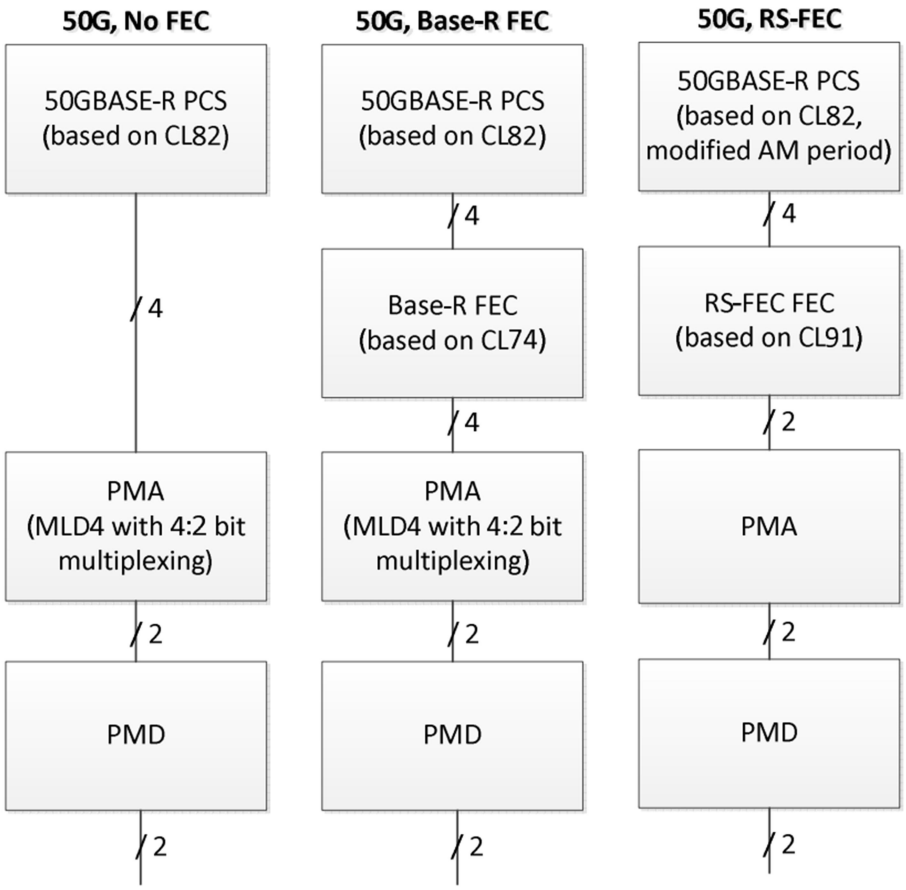
Řídicí signál je označován jako TXC (Transmit – vysílací) nebo RXC (Recieve – příjmací). Jeho velikost závisí na počtu 8b datových bloků. Na každý blok přichází jeden bit řídicího signálu. Tento signál udává zda se v daném bloku nachází data, bit je v logické ’0’, nebo kontrolní příkaz (začátek rámce, konec rámce, preambule, a další...) v tomto případě je bit nastaven do logické ’1’.

Poté co PCS podvrstva dostane data z XLGMII sběrnice (64b), proběhne kódování ze kterého získáme zakódovaný blok o velikosti 64b ke které je přidána 2-bitová synchronizační hlavička (hodnota "01"pro datové bloky a "10"pro kontrolní bloky). Následně je těchto 8B (bez synchronizační hlavičky) posláno do scrambleru, který s daty provede matematickou funkcí popsanou v klauzuli 49 ve standardu IEEE 802.3. blok scrambleru je možné přeskočit je-li potřeba (např. pro zrychlení simulace). Diagram funkce scrambleru znázorňuje Obr. 2.2.



Obr. 2.2: Diagram matematické funkce scrambleru [1]

Dále je k datům zpět přídána synchronizační hlavička. V dalším kroku je zapotřebí přidat zarovnávací značky, které se vkládají každých 16383 bloků. V případě 40G ethernetu, kdy PCS vrstva vysílá po čtyřech paralelních linkách (každá o šířce 64b + 2b synchronizační hlavičky) se vkládají celkem 4 zarovnávací značky (pro 100G ethernet pak 10 těchto značek). Jedna zarovnávací značka zabírá jeden celý 66-bitový blok (včetně 2b hlavičky). Hodnoty značek jsou dány tabulkou pro jednotlivou linku. První značka je vždy zarovnána na první linku. Posledním úkolem PCS podvrtvy je rozdělení jednotlivých bloků do vysílacích linek, a to kruhovým způsobem, tedy 1 -> 2 -> 3 -> 4 -> 1 -> ... Z těchto linek pak data pokračují do další podvrstvy, kterou je pro 50G ethernet buď podvrstva FEC (případně RS-FEC), nebo podvrstva PMA. Volba závisí na režimu, neboli módu, ve kterém pracuje fyzická vrstva. Rozdíl mezi těmito režimy znázorňuje Obr. 2.3.



Obr. 2.3: Rozdíl režimů pro fyzickou vrstvu [3]

2.3 FEC podvrstva

Podvrstva FEC je nepovinným prvkem pro fyzickou vrstvu, avšak některé režimy ethernetu jej vyžadují. Úlohou FEC je zkontrolovat odesílaný rámec a doplnit jej o kontrolní bity, které poté pomáhají příjmači odhalit chyby a opravit je. Jedná se o matematickou operaci jenž je velmi náročná na implementaci. Tato operace je opět popsána ve specifikaci od IEEE 802.3 v klauzuli 74.

2.4 PMA a PMD podvsrtvy

Tyto podvrstvy v zásadě implementuje sériový transciever z FPGA architektury. Tím tedy zajišťuje převod z 64B/66B podoby do sériových fyzických linek a naopak. Tyto linky pak vystupují pryč z FPGA a pokračují do FP (Form-factor Pluggable) transcieverů kde jsou vyvedeny do optických linek. Je-li potřeba, je zahrnut do PMA podvrstvy i bitový multiplex. Chování multiplexeru je dáno rovnicí popsanou opět ve standardu IEEE 802.3, jejíž parametry jsou počet vstupních a výstupních linek. Tyto podvrstvy mají také za úkol vyrovnat časové posuvy mezi linkami pro správnou práci s těmito daty.

3. HARDWAROVÉ SPECIFIKACE IMPLEMENTACE

3.1 Multi-Gigabitové Transcievery

MGT jsou obvodové prvky pracující na principu SerDes (Serializer/Deserializer) komunikace. Základní funkcí obvodu je převod sériového datového toku na paralelní pro příjmací část, a převod paralelního datového toku na sériový pro vysílací část. Typickým využitím gigabitových transcieverů jsou ethernetové systémy, bezdrátové routery, nebo například optické komunikační systémy.[2]

Pro FPGA obvody od firmy Xilinx jsou použity různé typy transcieverů pro různé architektury. Rozdíly jsou především v maximální rychlosti přenosu po jednotlivých linkách, pohybujících se v rozmezí od 3,2 Gb/s (GTP transcievery pro Spartan-6) až 58 Gb/s (GTM transcievery pro UltraScale+). Mezi další odlišnosti, dány hlavně vývojem, GT transcieverů od Xilinx patří například maximální počet možných transcieverů na architekturu, vnitřní struktura transcieverů nebo také více či méně podpůrných signálů.

Obrovskou výhodou GT transcieverů je především jednoduchá nastavitelnost a modifikovatelnost jednotlivých parametrů, jako jsou např. šířka vstupní a výstupní sběrnice, typ kódování (8B/10B, 64B/66B, 64B/67B), počet vstupních a výstupních portů, nebo již zmíněná rychlost přenosu po jedné lince.

3.2 FP transcievery

FP, neboli také Form-factor Pluggable transcievery, jsou síťové prvky sloužící pro příjem a vysílání datového toku primárně po optických komunikačních sítích. Existuje mnoho typů transcieverů, avšak na cílových kartách, na kterých bude prováděna implementace fyzické vrstvy, jsou použity konkrétně typy QSFP28 (Quad Small Form-factor Pluggable) pro karty NFB-200G2QL a NFB-100G2Q, a CFP4 (C Form-factor Pluggable) pro kartu NFB-100G2C. Použité typů FP transcieverů jsou znázorněny na Obr. 3.1.

QSFP28

Transcievery QSFP28 byly vytvořeny za účelem podpory 100G ethernetových aplikací. Mají k dispozici až 4 vysokorychlostní kanály, přičemž každý kanál je schopen přenášet rychlostí až 28 Gb/s (potenciálně až 40 Gb/s). Na kartách NFB-200G2QL a NFB-100G2Q se nacházejí vždy dva QSFP28 transcievery, ovšem v případě karty NFB-100G2Q je pouze jeden z nich kompatibilní s přenosovou rychlostí 25 Gb/s. Z tohoto důvodu může být pro standard 50GBASE-R použit pouze jeden z transcieverů.

CFP4

Transcievery CFP4 se nacházejí na kartě s označením NFB-100G2C. Podobně jako transcievery QSFP28, i tyto obsahují 4 kanály s rychlostí přenosu až 25 Gb/s, což splňuje předpoklady jak pro 100G standard, tak pro naši aplikaci 50G ethernetu.

3.3 Cílové síťové karty

Implementace fyzické vrstvy 50G ethernetu je kompatibilní a použitelná pro celkem 3 karty od firmy Netcope, a to konkrétně typy: NFB-200GQL, NFB-100G2Q a NFB-100G2C. Porovnání jednotlivých karet je uvedeno v Tab. 3.1. Karty jsou pak zobrazeny na Obr. 3.2.

Tab. 3.1: Porovnání specifikací FPGA karet od firmy Netcope

Typ karty	Transcievery	FPGA čip	Aktuálně podporované standardy
NFB-200GQL	2x QSFP28	Xilinx Virtex Ultrascale+	100GBASE-SR4/LR4 25G/50GBASE-SR 40GBASE-SR4/LR4 10GBASE-SR/LR
NFB-100G2Q	2x QSFP28	Xilinx Virtex-7 HT	100GBASE-LR4 40GBASE-SR4/LR4 10GBASE-SR/LR
NFB-100G2C	2x CFP4	Xilinx Virtex-7 XT	100GBASE-LR4 40GBASE-SR4/LR4 10GBASE-SR/LR

Obr. 3.1: FPGA karty pro implementaci 50G ethernetu [4]

4. NÁVRH FYZICKÉ VRSTVY

Následující kapitola se věnuje problematice návrhu fyzické vrstvy pro popis v jazyce VHDL a konečnou implementaci do obvodů do FPGA. Součástí návrhu je rozbor obou transcieverů GTY (UltraScale+) a GTZ (Virtex 7) a jejich generování včetně potřebných parametrů a odlišností.

4.1 Parametry návrhu

Není-li uvedeno jinak, platí všechny parametry obecně pro všechny karty a oba transcievery GTY a GTZ. Režim fyzické vrstvy pro implementaci je režim bez podvrstvy FEC, jednoduché blokové schéma je znázorněno na Obr. 2.3. Jedná se tedy o podvrstvy PCS, PMA a PMD. Pro plné využití potenciálu síťových karet, byl návrh vytvořen pro dva 50G vstupy. Popis ve VHDL byl psán genericky, kdy je možné pomocí parametrů zvolit cílovou architekturu (Virtex 7 nebo UltraScale+), a zda-li má být implementována jednoportová nebo dvouportová verze fyzické vrstvy.

4.2 PCS podvrstva

PCS podvrstva pro 50G ethernet má podle specifikace od Consortia stejnou funkcionalitu, vnitřní strukturu a parametry stejné jako pro 40G ethernet. Je tedy pro tento návrh použita již hotová implementace této podvrstvy.

PCS vrstva pracuje se dvěmi sběrnicemi. XLGMII, která slouží pro komunikaci mezi fyzickou a linkovou vrstvou. Obsahuje 3 signály pro každý směr (vysílací – TX, příjmací – RX). Datový signál o šířce 256b (TXD/RXD), kontrolní signál s šířkou 32b (TXC/RXC) a hodinový signál sběrnice (TXCLK/RXCLK). O hodinový signál XLGMII sběrnice se stará PMA podvrstva, konkrétně GT transciever.

Z druhé strany komunikuje PCS vrstva s podvrstvou PMA skrze 4 hlavní datové linky pro každý směr. Každá linka má šířku 66b, celkem tedy 4 x 66b mezi PCS a PMA podvrstvou. Po jedné lince se každý hodinový signál vysílá jeden 66b blok dat (kódování 64/66B), který obsahuje synchronizační hlavičku 2b a samotná data 64b. Hodinový signál pro část a přenos mezi PCS a PMA vrstvou se opět stará GT transciever, jenž je součástí PMA podvrstvy.

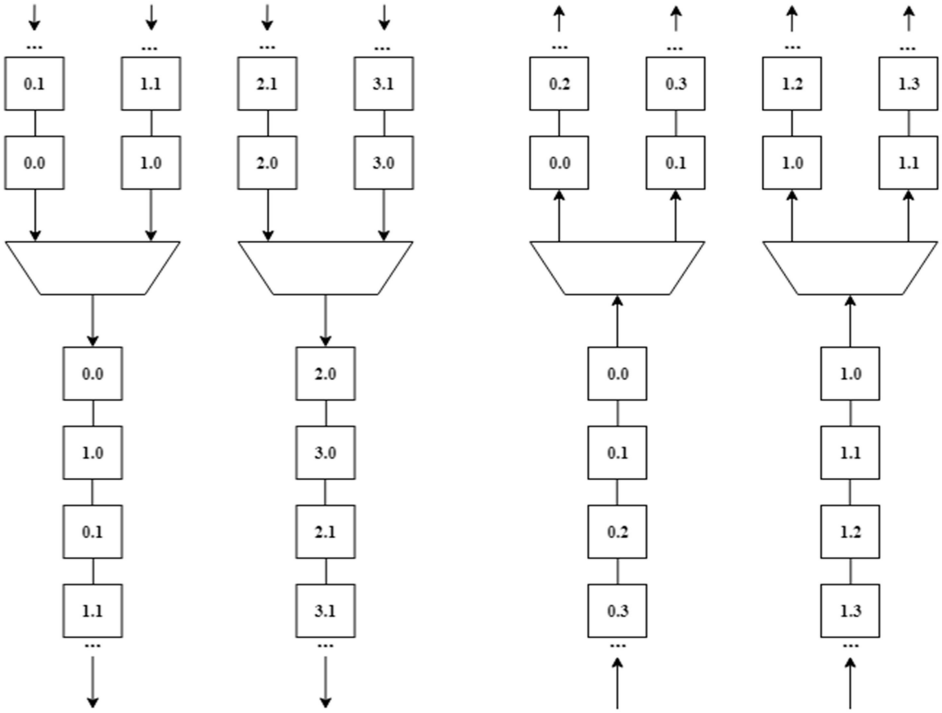
4.3 PMA a PMD podvrstva

PMA a PMD podvrstvy jsou součástí návrhu jako jeden blok. Tento blok se stará o veškerou komunikaci s deskou v rámci fyzické vrstvy a převod sériového datového toku na 66b datové bloky.

Zprostředkovává hodinový signál pro fyzickou vrstvu. Tento hodinový signál je fyzickým externím vstupem pro GT transciever, jehož frekvence je dána konstrukcí desky a je rovna hodnotě 322,265625 MHz.

PMA a PMD podvrstva vysílá a příjmá po sériových linkách, které jsou fyzicky propojeny s FP transcievery. Celkem se jedná o 8 sériových portů, 4 pro každý směr (RXN/TXN a RXP/TXP). Jedná se vždy o dvojici diferenciálních párů. Přenos po jedné lince probíhá rychlostí 25,78125 Gb/s. Ve výsledku tedy požadovaných 50 Gb/s.

Další důležitou funkcí PMA podvrstvy je v 50G režimu implementace bitového multiplexu 4:2 a 2:4. Kdy 4 datové linky z PCS podvrstvy multiplexuje do 2 virtuálních linek, které následně pokračují do GT transcieverů, a naopak. Funkci bitového mupltiplexu 4:2 a 2:4 znázorňuje Obr. 4.1 (Formát značení: k.n, kde k je číslo signálu a n je pořadí bitu).



Obr. 4.1: Funkce multiplexeru 4:2 a 2:4

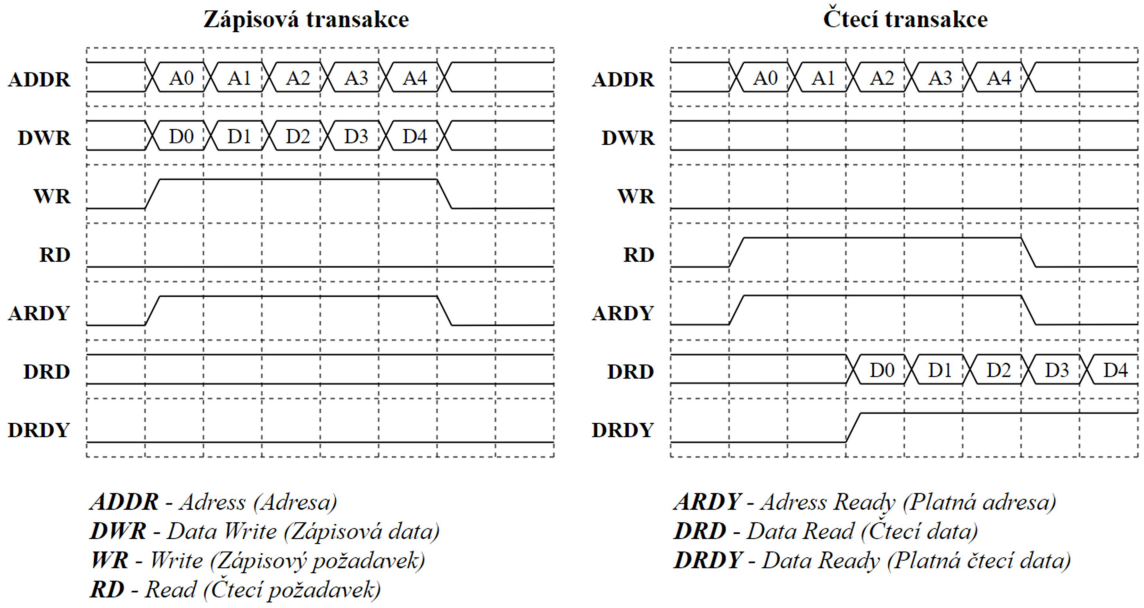
4.4 Řídicí blok

Posledním blokem návrhu fyzické vrstvy je řídicí blok (management). Jedná se o blok který zprostředkovává řízení fyzické vrstvy. Stará se o resetování PCS a PMA podvrstev, řízení podpurných signálů jako jsou například "LOOPBACK", který vytvoří zpětnou smyčku v PMA vrstvě (pomáhá při simulaci), nebo také "POWERDOWN", který při logické ’1’ vypne příslušnou linku z GT transcieveru. Součástí tohoto bloku je jsou také pomocné čítače stavů fyzické vrstvy jako chybovost, stav linek apod.

Tento blok funguje jako adresový dekodér který komunikuje přes MI32 rozhraní. Toto rozhraní obsahuje signály pro zápis a čtení, adresové a datové signály, a hodinový signál, jenž je odlisný od hodinových signálů použitých pro podvrstvy PCS a PMA.

MI32 sběrnice

MI32 sběrnice, nebo také "Memory Interface", slouží jako komunikační rozhraní s řídicím blokem. Zjednodušeně se jedná o paměť jejíž položky je možné přepisovat a číst z nich. Výjimkou jsou některé adresy, které jsou pouze pro čtení, a slouží jako informační (např. aktutálně nastavená rychlost karty, verze ethernetu apod.). V jednu chvíli lze buď zapisovat nebo číst. Příklad zápisové a čtecí transakce je znázorněn na Obr. 4.2.



Obr. 4.2: Příklad zápisové a čtecí transakce MI32 sběrnice

5. KOMPONENTY

5.1 Top modul

Pro přehlednost a snažší manipulaci s komponentou jako celek, byl top modul rozdělen do dvou souborů, entity a architektury. Entita ***50ge_phy_ent.vhd*** definuje rozhraní a generické parametry celého modulu. Zatímco architektura ***50ge_phy_arch.vhd*** obsahuje samotný kód a další podřazené komponenty.

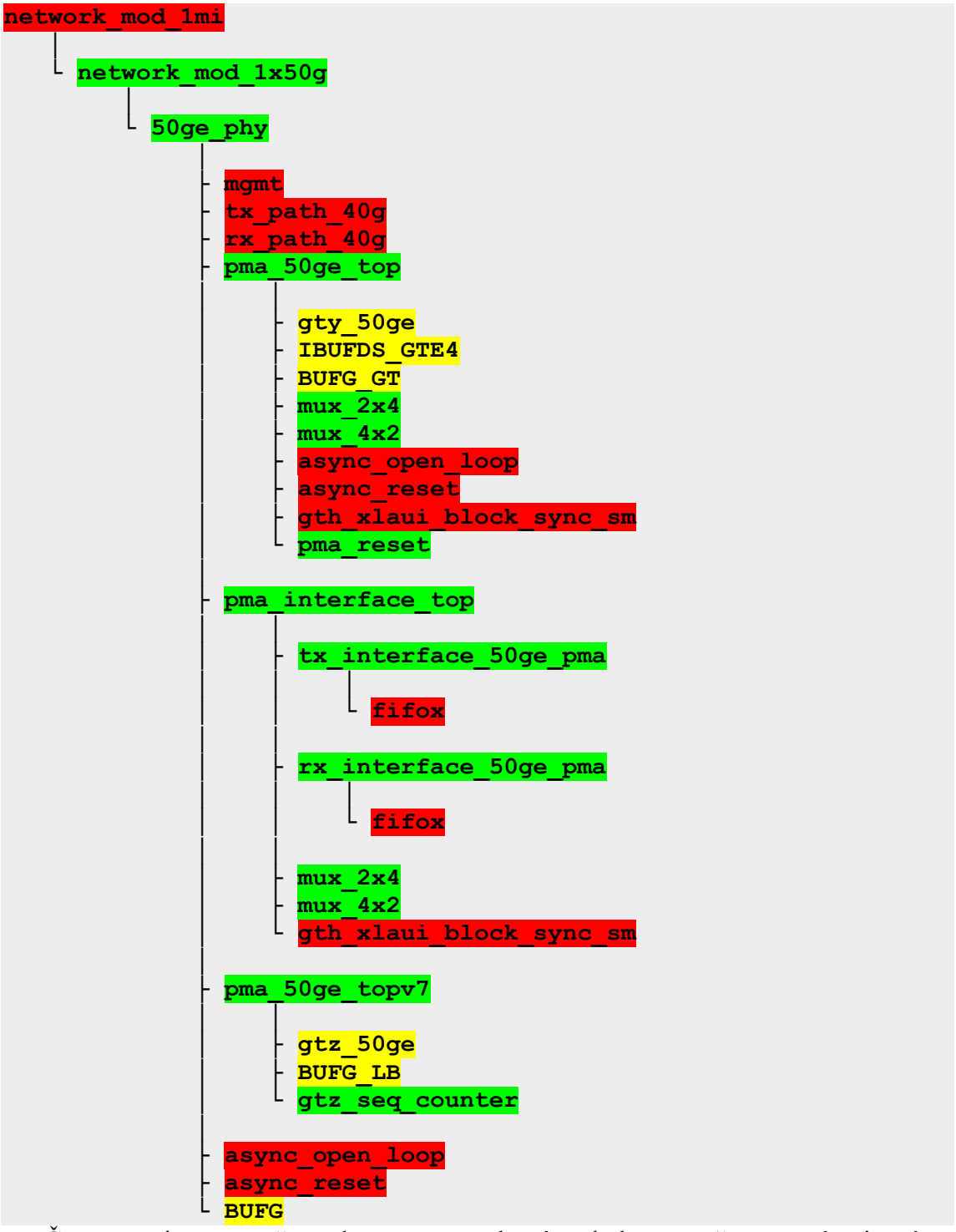
Top modul obsahuje tři generické parametry, DEVICE, PHY_NUMBER a mapování kanálů. Parametr DEVICE slouží ke zvolení typu cílové architektury FPGA. Buďto architekturu UltraScale+ "ULATRASCALE", nebo architekturu Virtex 7 "7SERIES". Jediným zásadním rozdílem při použití jiné architektury je v případě fyzické vrstvy pouze v integrovaném GT transcieveru (GTY nebo GTZ).

Parametr PHY_NUMBER pak určuje počet vstupních a výstupních 50G linek. Přesněji řečeno, při nastavení parametru na hodnotu 1, bude fyzická vrstva obsahovat jeden 50G port se dvěma 25G linkami. V případě nastavené hodnoty 2, bude fyzická vrstva pracovat se dvěma 50G porty, tedy 2x50G, neboli 4x25G linek.

Mapování kanálů, tedy generiky CH0_MAP až CH3_MAP, slouží ke správné organizaci (mapování) GT linek transcieverů. Vzhledem k tomu, že ne vždy mohou sériové linky vstupující do GT transcieverů odpovídat mapování na vstupu optických transcieverů, je třeba správně seřadit vstupy a výstupy GT transcieverů. V našem případě je prohozena třetí a první linka tedy posloupnost 3 – 1 – 2 – 0.

Tento modul je integrován do obálky pro 50G ethernet jako ***network_mod_1x50g*** a ta následně přidána do nadřazené komponenty ***network_mod_1mi***, která je pak integrována do konkrétní ethernetové karty.

Strom použitých komponent od nejvyšší vrstvy po nejnižší (komponenty, jež nejsou součástí práce, nejsou dále rozvíjeny):



Červenou jsou označeny komponenty, které nebyly vytvořeny v rámci práce. Zelenou jsou označeny komponenty, jež byly vytvořeny v rámci práce. Žlutá barva značí komponenty, které jsou součástí FPGA knihoven, nebo byly vytvořeny při generování pomocí prostředí Vivado.

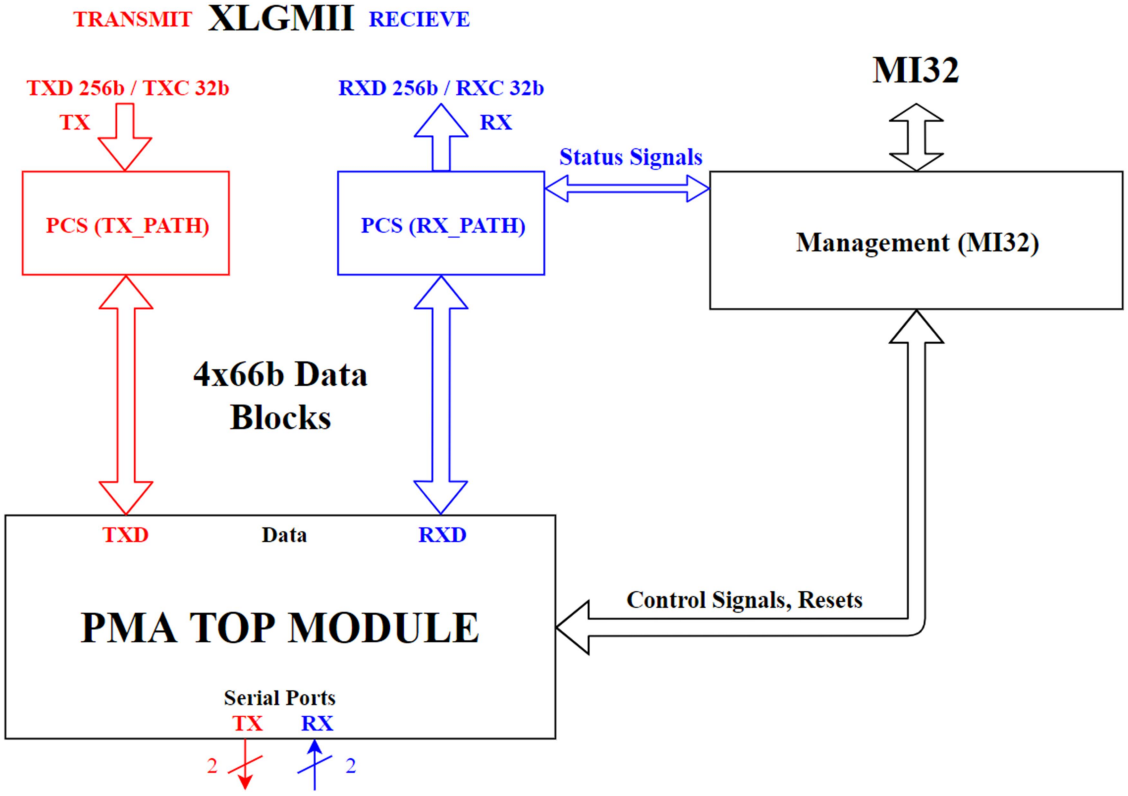
Porty rozhraní fyzické vrstvy

První skupinou portů jsou hodinové a resetovací signály. Singál RESET slouží jako globální reset pro celou komponentu. CLK_STABLE určuje zda-li je fyzická vrstva připravena k činnosti a má stabilní hodinové signály. REFCLK_OUT je hodinový signál, který generuje příslušný GT transciever, Podle požadavků mají tyto hodiny frekvenci 322,265625 MHz. DRPCLK jsou hodiny pro GTY transciever, a slouží jako volně běžící hodiny potřebné k resetování základních stavů tzv. primitivů. Tento hodinový signál je napřímo připojen k vstupu GTY transcieveru gtwiz_reset_clk_freerun_in. Hodinové singály REFCLK_P a REFCLK_N slouží ke generování referenčních hodin REFCLK_OUT. Tento hodinový signál má frekvenci 156,25 MHz.

Další skupinou portů jsou rozhraní XLGMII sběrnice. Hodinové signály RXCLK a TXCLK vycházejí ze stejných hodin jako REFCLK_OUT, avšak pro tyto hodiny je použit navíc buffer ***BUFG***. Zbylé porty jsou datové xxXD a řídící xxXC signály pro XLGMII sběrnici.

Pro řízení řídicího bloku tzv. managementu, jsou vyhrazeny porty MI_xx. Funkce jednotlivých signálů je popsána v kapitole 4.4. Hodinový signál MI_CLK je zprostředkován z nadřazené komponenty.

Jako poslední důležitou skupinou portů je sériové rozhraní fyzické vrstvy. Seriové porty pro vysílací a příjimací část jsou tvořený jako diferenciální pár. Polaritu těchto portů lze zaměnit signály RXPOLARITY a TXPOLARITY.



Obr. 5.1: Blokové schéma top modulu fyzické vrstvy

Architektura top modulu

Hlavním obsahem architektury top modulu je instancování jednotlivých použitých komponent. Na Obr. 5.1 je zobrazeno stručné blokové schéma top modulu fyzické vrstvy. Nejsou zde zobrazeny všechny komponenty a signály s menší prioritou. Červenou barvou je označena vysílací cesta, modrou pak přijímací. Seznam použitých komponent, včetně informace zda-li je její popis součástí bakalářské práce, znázorňuje Tab. 5.1.

Krom instancování komponent je součástí top modulu také podpůrná logika řídící propojení a správnou funkčnost těchto komponent. Mezi tuto logiku se řadí všechny asynchronní přechody, řízení resetovacích signálů a generování hodin pro XLGMII sběrnici pomocí bufferu BUFG.

5.2 Management (Řídicí blok)

Hlavní účel této komponenty ***mgmt.vhd*** je popsán již v kapitole 4.4. Tato komponenta je společná pro všechny rychlosti ethernetu a jak o celek je převzatá a instancována do top modulu fyzické vrstvy.

Slouží hlavně pro řízení resetovacích signálů a kontrolu stavu podpůrných signálů. Jedná se převážné o stavové a ladící signály přicházející z příjmací části PCS vrstvy, a to konkrétně komponenty ***rx_path_40g***. Jsou to například informace o stavech podpůrných čítačů, PCS a PMA linek apod. Tyto informace slouží zejména při ladění fyzické vrstvy, kdy si stavy těchto čítačů a signálů můžeme v reálnem čase nechat vypsat do konzole který komunikuje přímo s kartou.

5.3 PCS vrstva

O realizaci PCS vsrstvy v tomto návrhu se starají již dokončené popisy převzaté z verze ethernetu pro rychlost 40G (vychází se z definice 50G ethernetu). Jedná se o popisy ***rx_path_40g*** pro přijímací část a ***tx_path_40g*** pro vysílací část. Funkci PCS vrstvy je věnována samostatná kapitola 2.2. Popis těchto komponent není součástí bakalářské práce.

Tab. 5.1: Seznam použitých a vytvořených komponent

Název komponenty	Název souboru	Vlastní popis
MI32 Management	*mgmt.vhd*	Ne
PCS TX Path	*tx_path_40g.vhd*	Ne
PCS RX Path	*rx_path_40g.vhd*	Ne
PMA GTY Module	*pma_50ge_top.vhd*	**Ano**
PMA GTZ Interface	*pma_interface_top.vhd*	**Ano**
PMA GTZ Module	*pma_50ge_topv7.vhd*	**Ano**
Asynchronous Open Loop	*async_open_loop.vhd*	Ne
Asynchronous Reset	*async_reset.vhd*	Ne
Global Clock Simple Buffer	*BUFG*	FPGA knihovna
Bit Multiplexer 4:2	*mux_4x2.vhd*	**Ano**
Bit Multiplexer 2:4	*mux_2x4.vhd*	**Ano**
Gigabit Transceiver Buffer [9]	*IBUFDS_GTE4*	FPGA knihovna
Clock Buffer Driven by GTY [9]	*BUFG_GT*	FPGA knihovna
Block Lock Sync	*gth_xlaui_block_sync_sm.vhd*	Ne
GTZ Sequence Counter	*gtz_seq_counter.vhd*	**Ano**
GTZ Loopback Clock Buffer [10]	*BUFG_LB*	FPGA knihovna
TX GTZ Interface Logic	*tx_interface_50ge_pma*	**Ano**
RX GTZ Interface Logic	*tx_interface_50ge_pma*	**Ano**
Universal FIFO	*fifox.vhd*	Ne
GTY PMA Reset Counter	*pma_reset.vhd*	**Ano**
GTY IP Core	*gty_50ge.vho*	IP jádro
GTZ IP Core	*gtz_50ge.vho*	IP jádro
Network Module 50G	*network_mod_1x50g.vhd*	**Ano**
Network Module Top	*network_mod_1mi.vhd*	Ne

5.4 PMA Vrstva pro UltraScale+

PMA vrstva je rozhraním mezi fyzickou sériovou linkou a PCS vrstvou. Hlavním úkolem této vrstvy je implementovat gigabitový transciever, který se liší podle použité architektury FPGA.

Generování GTY transcieveru

Před samotnou implementací a instancováním gigabitových transcieverů je zapotřebí tyto transcievery vygenerovat jako IP (Intellectual Property) jádro s požadovanými vlastnostmi a vstupními a výstupními porty. Generování GTZ transcieveru bylo provedeno přes UltraScale FPGAs Transcievers Wizard (1.7). Po vygenerování a syntetizování transcieveru se vedle hlavních konfiguračních souborů vygeneruje take kód pro instancování konkrétní vygenerované verze, kterou je zapotřebí instancovat jako komponentu v top modulu PMA vrstvy (***pma_50ge_top***).

Důležité požadavky a nastavené parametry transcieveru:

- Line Rate = 25,78125 Gb/s (pro oba směry RX/TX)
- Reference Clock = 322,265625 MHz (pro oba směry RX/TX)
- Encoding = Async. gearbox for 64/66B (pro oba směry RX/TX)
- Used data width = 128b
- Internal data width = 64b
- Free-running Clock = 195,3125 MHz
- Počet kanálů = 4

Pro další funkce a řízení transcieveru byly přidány následující porty:

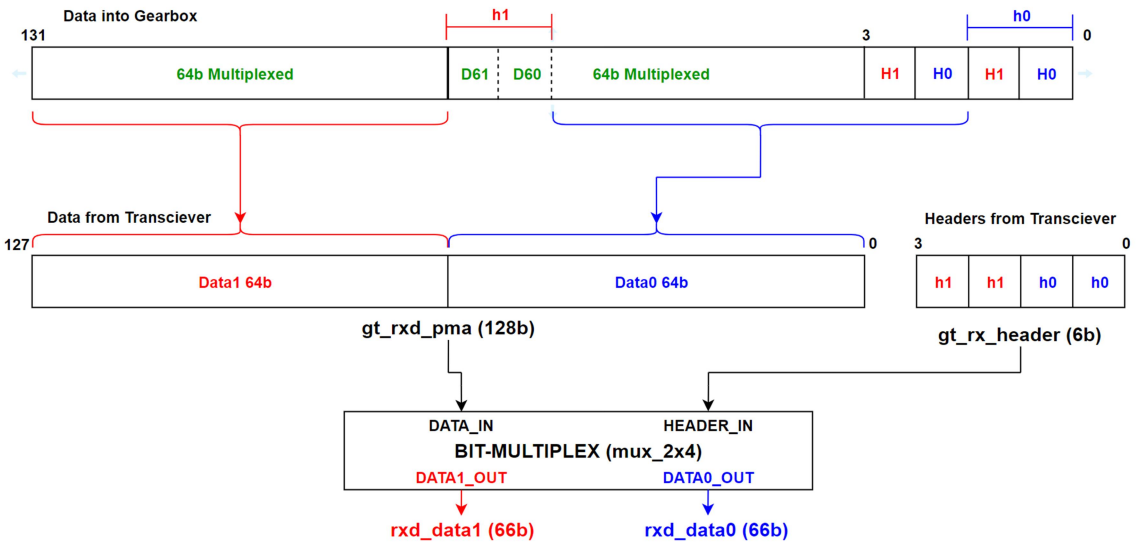
- loopback_in – Nastavením do jedničky vytvoří zpětnou smyčku dat, slouží pouze pro ladící účely.
- rxpd_in a txpd_in – Porty pro "uspání" linky transcieveru. Nemá zásadní význam pro funkčnost.
- rxpmareset_in – Slouží pro resetování vysílacích linek. Pravidelný reset není-li linka správně chycena (nejsou aktivní block locky).
- rxpolarity_in a txpolarity_in – Prohodí polaritu sériových linek. Aktivní pouze je-li to vyžadováno.

Zpracování dat a bitový multiplex

Vzhledem k tomu, že v případě 50G ethernetu jsou data včetně hlaviček multiplexovány do sebe, tedy po jedné lince ve 132 bitech jsou vysílány dva 66 bitové datové bloky včetně hlaviček. Je tedy zapotřebí tyto data opět rozdělit a vyslat po 2 virtuálních linkách dále do PCS vrstvy. Stejný princip platí pro vysílací TX směr, kdy ze dvou 66 bitových bloků je nutné vytvořit 132 bitový blok a ten poslat do GTY transcieveru.

GTY transciever je nastaven tak, že na datovém rozhraní je datová šířka 128 bitů (User data width = 128) a data jsou zpracovavány asynchronním gearboxem s 64/66B enkódovaním (Encoding = Async. gearbox for 64/66B). Na výstupu transcieveru je tedy k dispozici 128 bitů dat + příslušných 6 bitů hlaviček (bity 3 a 6 jsou nulové a v tomto režimu nevyužity). Avšak před samotným bitovým multiplexem je třeba data a hlavičky správně uspořádat do 132 bitového bloku. Asynchronní gearbox 64/66B rozděluje data a hlavičky na výstup tak, jak kdyby data chodila již v 66 bitových blocích. Proces rozdělení je zobrazen na Obr 5.2.

Tento proces je opakován pro každou linku. V případě čtyř PMA linek získáme ve výsledku osm linek pro PCS vrstvu. Obdobným způsoben je řešena vysílací cesta, kdy je třeba z multiplexovaného 132b bloku získat 128b + 6b pro vstup transcieveru s tím, že transciever s těmito daty počítá jako se dvěma 66b bloky.



Obr. 5.2: Proces rozdělení výstupních dat GTY transcieveru na 66 bitové bloky

Implementace block lock logiky

Dalším velmi podstatným prvkem PMA vrstvy je implementace a řízení block lock logiky. Tuto funkci zaopatřuje komponenta ***gth_xlaui_BLOCK_SYNC_SM***. Block lock je signál, jehož hodnota určuje, zda-li linka GTY posílá správně zarovnaná data. Tato logika je řízena datovými hlavičkami vysílanými z GTY transcieveru. Block lock (BLOCKSYNC_OUT) se nastaví do logické '1' je-li příjmaná hlavička validní po definovaný počet taktů (SH_CNT_MAX = 1024). Není-li hlavička validní po dobu 64 taktů (SH_INVALID_CNT_MAX = 64) pak komponenta nastaví výstupní signál RXGEARBOXSLIP_OUT do logické '1'. Tímto signálem je řízen GTY port rxgearboxslip_in. Je-li port nastaven do logické '1', asynchronní gearbox posune výstupní data o jednu pozici (jeden bit). Tento proces se opakuje do doby, než se správně zarovnají data a nastaví block lock to '1'.

Block locky se vypočítávají pro každou PCS linku zvlášť, tedy celkem 8 block locků pro 4 PMA linky. Vzhledem k tomu, že každá PMA linka má dvě PCS linky, tedy i 2 block lock a gearboxslip signály, je nutné vybrat který gearboxslip bude vstupovat to transcieveru (pouze jeden gearboxslip_in pro jednu PMA linku). Není-li locknutá (block_lock = '0') první PCS linka, do transcieveru vstupuje první gearboxslip. Je-li aktivní první, ale ne druhá, vstupuje do transcieveru druhý gearboxslip. Tímto se zamezí problému když by se chytila pouze jedna z linek.

Dále je zapotřebí ošetřit stav, který by mohl nastat při odpojení linky. V tomto případě je třeba resetovat přijímací stranu RX transcieveru. To je vyřešeno čítačem který při neaktivních block lock signálech resetuje PMA linky přibližně každých 100 ms, dokud se linka nezachytí (všechny block locky aktivní).

Řízení hodinových signálů

GTY transciever má několik vstupních hodinových portů, které je třeba řídit podle příslušné specifikace daného transcieveru.

Prvním podstatným hodinovým signálem jsou referenční hodiny pro fázový závěs PLL (Phase-Locked Loop). Tyto hodiny je podle specifikace generovat přes vnitřní buffer GT transcieveru, konkrétně IBUFDS_GTE4, kdy vstupní reference jsou hodinové signály REFCLK_P_IN a REFCLK_N_IN.

Pro funkci transcieveru je vyžadován také jeden volně bežící hodinový signál u kterého je zajištěno, že je stabilní a beží po celou dobu. K těmto účelům slouží DRPCLK hodiny s frekvencí 125 MHz. Tyto hodiny vedou na port transcieveru gtwiz_reset_clk_freerun_in.

Nejpodstatnější skupinou hodinových signálů jsou hodiny generovány z výstupů transcieveru txoutclk_out a rxoutclk_out (pro každou linku jeden hodinový výstup). Podle specifikací musí tyto výstupy využívat hodinového bufferu BUFG_GT řízeného GT transcieverem. Výstupem těchto bufferů jsou pak hodinové signály sloužící k řízení veškeré logiky uvnitř a vně transcieveru. Těmito signály jsou rxusrclk_in a txusrclk_in, a rxusrclk2_in a txusrclk2_in. Frekvence těchto hodin je definována podle specifikací, kdy rozhodujícími parametry jsou interní a externí datové šířky. V případě 128b externí a 64b interní šířky je definováno:

$$F_{RXUSRCLK} = F_{RXOUTCLK}$$
$$F_{TXUSRCLK} = F_{TXOUTCLK}$$
$$F_{RXUSRCLK2} = F_{RXUSRCLK}/2$$
$$F_{TXUSRCLK2} = F_{TXUSRCLK}/2$$

Toto dělení zajišťuje BUFG_GT, který má nastavitelný vstup pro velikost podílu vstupního a výstupního signálu. Hodiny txusrclk_in jsou mimo jiné využity pro nastavení portu gtwiz_userclk_rx_active_in, který je nutné nastavit do logické ‘1’ jakmile jsou hodiny txusrclk_in aktivní pro funkčnost příslušných pomocných bloků transcieveru. Totéž platí pro hodinový signál rxusrclk_in a port gtwiz_userclk_rx_active_in.

Hodinovými signály rxusrclk2_in a txusrclk2_in je řízena veškerá logika kolem zpracování dat včetně PCS vrstvy. Pro vysílací TX část jsou využity pouze dva hodinové buffery BUFG_GT, jeden pro txusrclk_in a druhý pro txusrclk2_in. Tyto dva hodinové signály jsou použity pro všechny PCS a PMA linky. Avšak v případě příjimací RX části je pro každou PCS a PMA linku individuální BUFG_GT a tedy čtyři rxusrclk_in a čtyři rxusrclk2_in. Hodinovými signály rxusrclk2_in je mimo jiné řízena block lock logika a resetování PMA linky

Zbylá logika vně PMA vrstvy zahrnuje nastavení podpurných a ladících portů (loopback_in, txpd_in, rxpd_in) a stavových signálů, jejichž hodnota udává správnost dat a hlaviček (rxd_valid), a zda-li je dokončen reset transcieveru (pmarxreset_done a pmatxreset_done).

5.5 PMA vrstva pro Virtex 7

Popis PMA vrstvy pro GTZ transciever architektury Virtex 7 nebyl žádným způsobem testován. Popis této PMA vrstvy tedy není finální a jedná se pouze o předběžný návrh a popis. Nejsou tedy implementovány chyby nalezeny při testovaní implementace s GTY transcieverem, které jsou s největší pravděpodovností přítomny také u verze pro Virtex 7.

Jediným rozdílem mezi implementací pro Virtex 7 a UltraScale+ je v použitém transcieveru. GTZ transciever na rozdíl od transcieveru GTY jiným způsobem pracuje s výstupními a vstupními daty na přechodu PCS a PMA vrstvy. Dalším podstatným rozdílem je to, že transciever GTZ neumožňuje implementaci asynchronního gearboxu jakož tomu je transcieveru GTY. Z těchto důvodu bylo zapotřebí implementovat další komponenty, které řeší tyto odlišnosti.

Generování GTZ transcieveru

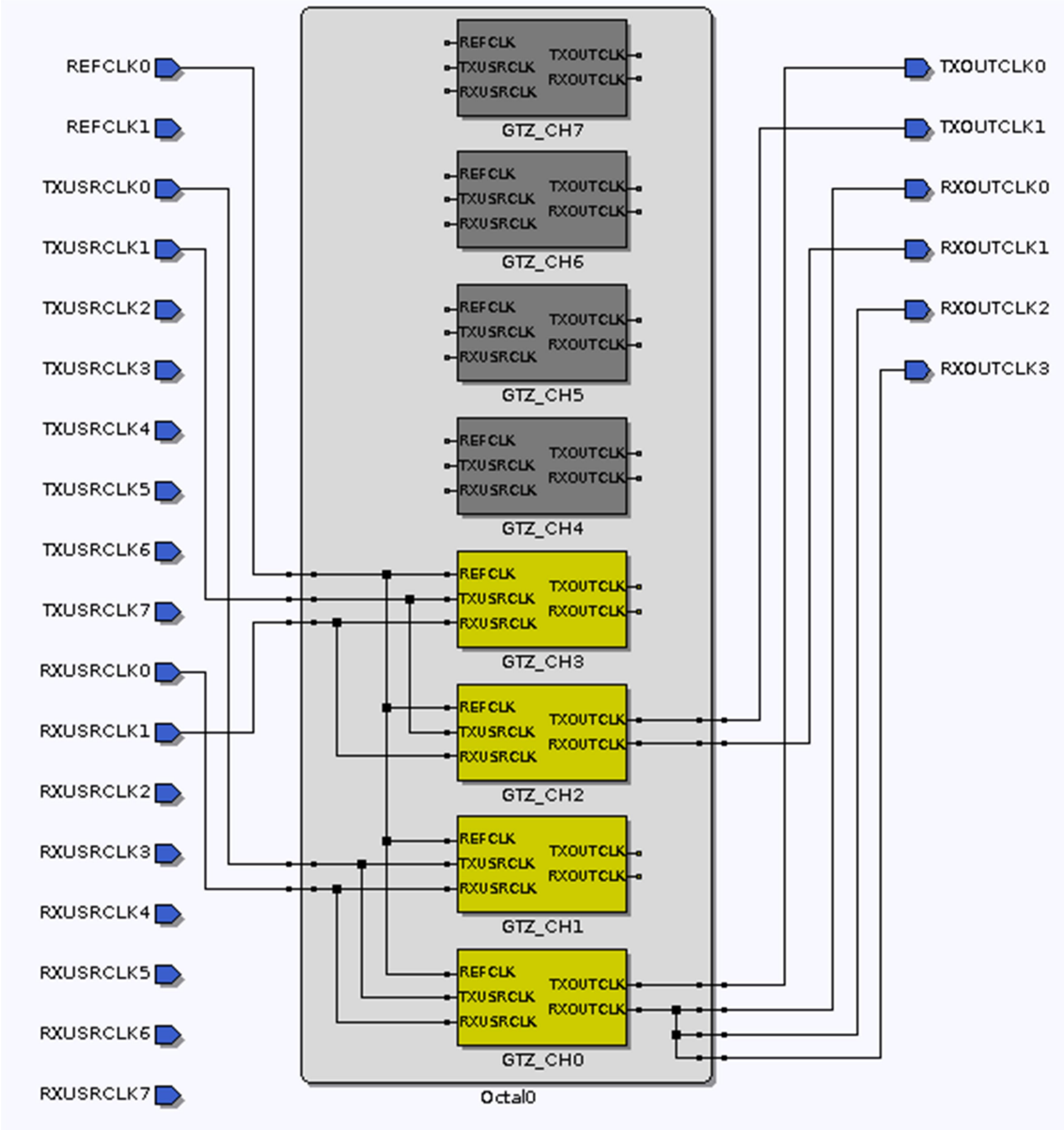
Generování GTZ transcieveru bylo provedeno přes 7 Series FPGAs Transcievers Wizard (3.6).

Důležité požadavky a nastavené parametry transcieveru:

- Line Rate = 25,78125 Gb/s (pro oba směry RX/TX)
- REFCLK0 Source = 322,266 MHz
- FIB mode (gearbox mode) = 64/66B
- Fabric data width = 160b (pro všechny 4 linky)
- USRCLK a OUTCLK = 161,133 MHz
- Počet kanálů = 4

Pro další funkce a řízení transcieveru byly přidány následující porty:

- rxen_in a txen_in – Porty pro "uspání" linky transcieveru. Nemá zásadní význam pro funkčnost. Stejný princip jako v porty rxpd_in a txpd_in u GTY transcieveru (negativní logika)
- rxpolarity_in a txpolarity_in – Prohodí polaritu sériových linek. Aktivní pouze je-li to vyžadováno.



Obr. 5.3: Zvolené kanály GTZ transcieveru včetně zdrojů hodinových signálů

První z těchto komponent je komponenta ***gtz_seq_counter***. Vzhledem k tomu, že se v případě GTZ transcieveru pracuje se synchronním gearboxem, bylo zapotřebí, podle specifikací pro GTZ transciever, implementovat čítač sekvencí, který s každým hodinovým cyklem (hodinový signál pro vysílací TX směr) přičte '1'. Výstupní hodnota čítače je přivedena na porty transcieveru txsequence0_in až txsequence3_in Dále čítač každých 32 hodinových cyklů nastaví výstupy DATAREADY a HEADREADY do log '0' (jinak '1') a resetuje celý čítač. Signály DATAREADY a HEADREADY určují zda-li je možné vysílat data do transcieveru.

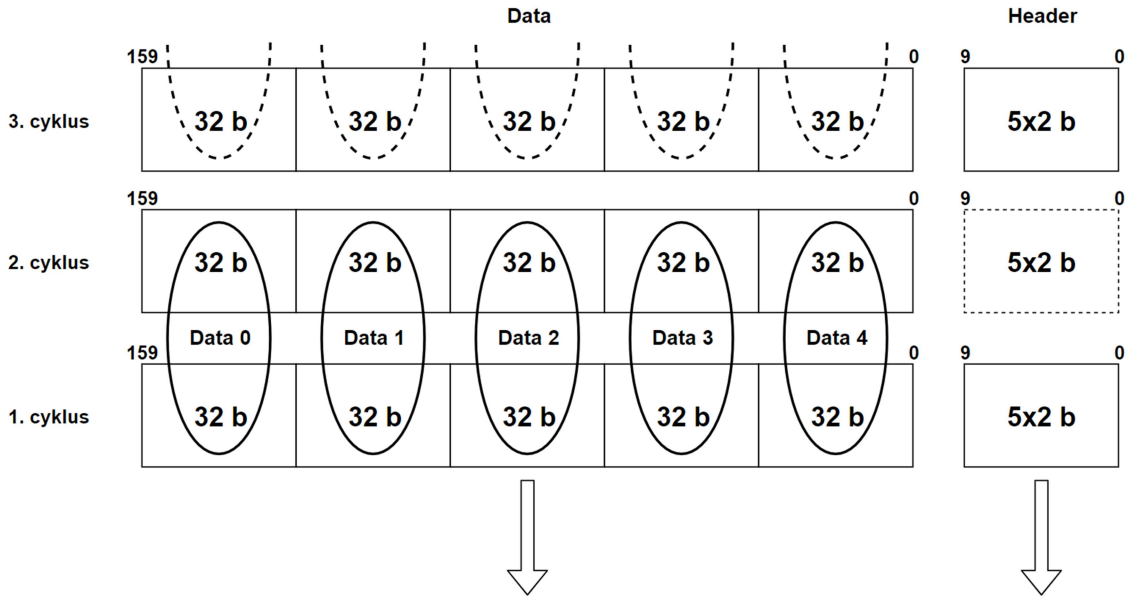
Řízení hodinových signálů

Hodinové porty jsou u GTZ transcieveru řešeny obdobně jako u verze GTY. Pro řízení hodin jsou využity GTZ buffery BUFG_LB s jedním hodinovým vstupem a jedním výstupem. Jako vstupy jsou využity výstupní porty z GTZ transcieveru txoutclk0_out, txoutclk1_out, rxoutclk0_out a rxoutclk1_out. Výstupy bufferů jsou přivedeny na porty transcieveru txusrclk0_in, txusrclk1_in, rxusrclk0_in a rxusrclk1_in respektive. Výstupy bufferů také slouží jako zdroj hodin zbylé logiky a PCS vrstvy, podobně jak je tomu v případě transcieveru GTY.

Jako zdroj hodin pro referenční hodinový port GTZ transcieveru jsou opět využity porty REFCLK_N_IN a REFCLK_P_IN. Avšak v tomto případě je implementace jednodušší, a tyto signály jsou připojeny přímo na porty gtrefclk0n_in a gtrefclk0p_in. Zdroje hodin pro zvolené kanály je možné vidět na Obr. 5.3.

Zpracování dat a bitový multiplex

Pro zpracovávání dat na rozhraní PMA a PCS vrstvy pro GTZ transciever bylo zapotřebí vytvořit komponentu, která by správně přeskládávala vstupní a výstupní data transcieveru. Transcieveru tyto data přeskládává podle systému zobrazeném na Obr. 5.4.



Obr. 5.4: Systém výstupních/vstupních dat GTZ transcieveru na rozhraní PCS a PMA vrstvy

Data jsou vysílány po 160 bitových rozhraních po 64b blocích tak, že první polovina 64b bloku (32b) je odeslána v prvním taktu, společně s příslušnými dvěma bity hlavičky a dalšími čtyřmi polovinami 64b bloků. Celkem se tedy v prvním taktu odešle 5x32b dat + 5x2b hlaviček. V druhém taktu vyšle transciever zbylou polovinu dat, tedy zbylých 32b. V tomto taktu je hlavičkový výstup nevyužit. Stejným způsobem je třeba vysílat data do transcieveru v TX směru. Po (RX směr) a před (TX směr) je třeba implementovat bitový multiplex jako v případě GTY transcieveru. K účelům zpracování těchto dat slouží komponenty ***pma_interface_top***, ***rx_interface_50ge_pma*** a ***tx_interace_50ge_pma***.

Komponenta ***pma_interface_top*** slouží pouze jako top modul pro implementaci jednotlivých rozhraní a bitového multiplexu. Zbylé dvě komponenty realizují zpracování těchto dat v RX (***rx_interface_50ge_pma***) a TX (***tx_interace_50ge_pma***) směru. Přeskládávání dat je realizováno pomocí registrů, které ukládají data z předchozího taktu a paměti FIFO, která ukládá již přeskládána data a vysílá je dál.

6. SYNTÉZA A TESTOVÁNÍ

6.1 Syntéza

Syntéza kódu a implementace kódu byla prováděna pomocí softwaru Vivado 2018.2. Použitým FPGA obvodem byl typ FLVB2104-XCVU7P.

Využití zdrojů

Tab. 6.1. obsahuje data a informace o využití zdrojů fyzické vrstvy pro jednoportovou verzi při použité architektuře UltraScale+.

Tab. 6.1: Využití zdrojů fyzické vrstvy pro jednoportovou verzi UltraScale+

Typ zdroje	Využití	Využití v procentech [%]
CLB LUT	10199	1,29
CLB Registers	16102	1,02
CARRY8	130	0,13
F7 Muxes	139	0,04
F8 Muxes	64	0,03
CLB	2786	2,83
LUT as Logic	8979	1,14
LUT as Memory	1220	0,31
LUT Flip Flop Pairs	4331	0,55
Block RAM Tile	37,5	2,60
Global Clock Buffers	10	0,83
GTYE4_CHANNEL	4	5,26
GTYE4_COMMON	1	5,26

Časová analýza

V Tab. 6.2 jsou uvedeny hodnoty časové analýzy výsledné implementace výsledné karty (nejedná se pouze o fyzickou vrstvu). Z těchto výsledků můžeme vidět, že implementace splňuje časové podmínky. Co se tedy týče časových podmínek, tak fyzická vrstva včetně celé implementace karty vyhovuje požadavkům.

Tab. 6.2: Výsledky časové analýzy implementace síťové karty

WNS (Worst Negative Slack) [ns]	0,032
TNS (Total Negative Slack) [ns]	0,000
WHS (Worst Hold Slack) [ns]	0,010
THS (Total Hold Slack) [ns]	0,000

6.2 Testování a ověření funkčnosti

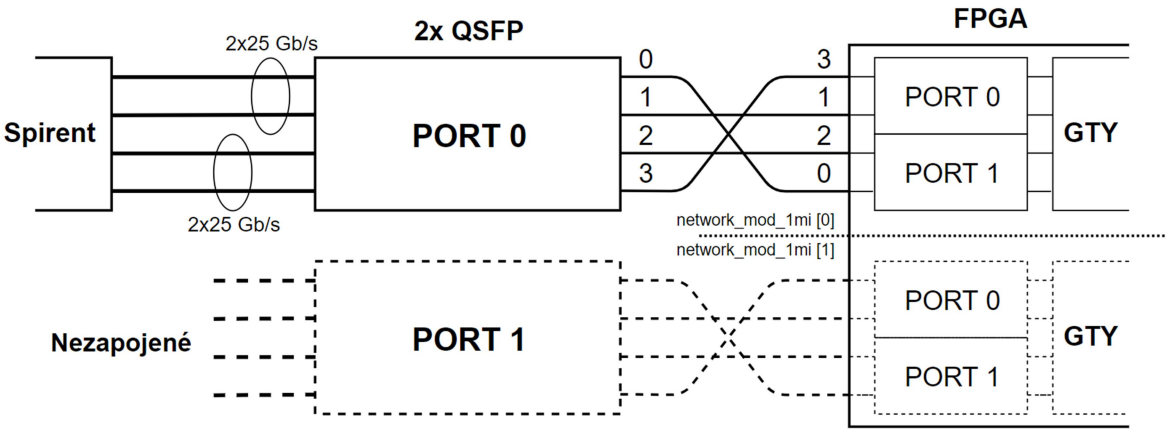
Jako prioritu mělo při testování verze fyzické vrstvy pro architekturu UltraScale+, a to vzhledem k tomu, že tato architektura je na rozdíl od Virtex 7 modernější a mnohem perspektivnější.

Před samotným testováním je zapotřebí implementovat popis fyzické vrstvy do již existujícího popisu akcelerační karty. Pro tyto účely je třeba vytvořit obálku, která fyzickou vrstvu instancuje a propojí s ostatními komponentami a fyzickými porty. O toto instancování se stará komponenta ***network_mod_1x50g***, která je dále instancována do vyšší vrstvy ***network_mod_1mi***.

Testování 50G ethernetu pro architekturu UltraScale+ bylo prováděno na kartě NFB-200G2QLP která obsahuje dva QSFP28 porty, FPGA s architekturou UltraScale+. Testování probíhalo pouze na prvním portu, který je připojen k testovacímu modulu Spirent FX3-100GTN-T2. Spirent je ovládán pomocí aplikace Spirent TestCenter kde je také možné nastavovat parametry portů jako jsou rychlost, nebo také režim testovaného ethernetu a ruzných klauzulí.

Porty ze Spirent modulu jsou pak propojeny s optickými transcievery testované karty NFB-200G2QLP. Tato karta je nainstalována přes PCI sběrnici k řídicímu serveru přes který komunikuje. Hlavním využitím serveru je nahrání bitových souborů a sledování stavu jednotlivých parametrů implementace. Například příkazem “nfb-eth”, který do konzole vypíše aktuální stav fyzické vrstvy, tedy zda-li jsou aktivní PMA a PCS linky, jsou-li aktivní block locky a další užitečné informace k odlazení implementace.

Propojení portů akcelerační karty a Spirent modulu je vidět na Obr. 6.1. Se Spirentem je propojen pouze jeden optický transciever. Pro jeden 50G port jsou použity dvě první linky, pro verzi 2x50G jsou pak využity všechny čtyři linky. Ve skutečnosti je toto řešeno tak, že transciever vždy zpracovává všechny čtyři linky, avšak generickým parametrem PHY_NUMBER se nastaví, zda-li top modul bude implementovat PCS vrstvu a logiku kolem pro všechny čtyři linky (2x50G) nebo pouze první dvě (1x50G).



Obr. 6.1: Propojení portů Spirent modulu a akcelerační karty

Simulace a verifikace obvodu

Běžným postupem při testování funkčnosti VHDL popisu je simulace a verifikace daného obvodu a logiky. Vytvoří se verifikační nebo simulační prostředí, které simuluje požadovaný průběh dat na rozhraní testovaného popisu, a sleduje se výsledná odezva, zda-li souhlasí s očekávaným výstupem.

Při testování implementace fyzické vrstvy 50G ethernetu jsme se rozhodli tento popis nesimulovat, ale testovat jej rovnou na akcelerační kartě v reálném chodu karty. Důvodem tohoto rozhodnutí bylo hlavně z časových důvodů. Samotná příprava simulačního prostředí a příslušných modelů, včetně testování průchodu dat, které by trvalo i několik hodin, by bylo z hlediska času neefektivní. Jelikož popis zahrnuje z větší části pouze instancování již hotových komponent, a samotnou logiku kolem, bylo jednodušší ladit a upravovat popis přímo na síťové kartě.

K ladícím účelům sloužila ILA (Integrated Logic Analyzer) sonda, která komunikuje s prostředím Vivado (Hardware Manager) skrze JTAG rozhraní FPGA obvodu, kde je sonda implementována. Sonda pracuje na daném vstupním hodinovém signálů a vzorkuje data přiváděné na vstup. Tyto data je pak možné vyčíst a zobrazit v Hardware Manageru v podobě průběhů v grafické podobě.

Testování přes Spirent TestCenter

Spirent pro účely testování nabízí několik funkcí. Nejpodstatnější pro nás jsou funkce generování rámců podle zvolených parametrů, měření rychlosti přenosu na TX i RX směru, čítače příchozích a odchozích rámců a jeho délek, a také zachytavání rácmů podle nastavených podmínek což slouží pro kontrolu chybnosti příchozích rámců. Takto zachycené rámce je možné zobrazit pomocí programu Wireshark, který vypíše všechny údaje o rámcích, včetně jejich obsahu.

Na Obr. 6.2 je vidět výstup z konzole, který ukazuje aktuální stav fyzické vrstvy po nahrání firmwaru do karty a nastavení správných parametrů na Spirent modulu. Výstupy do konzole jsou získávány přes MI32 sběrnici a Management modul instancovaný ve fyzické vrstvě. Důležitými informace jsou hlavně stavy linek (Link status) pro PCS a PMA vrstvu, stavy block a AM locků, čítače BIP chyb (BIP error counters) a chyby na linkách (Transmit/Recieve Fault). Vidíme, že všechny linky na PCS a PMA vrstvě jsou aktivní (“UP”). Tak stejně jsou aktivní všechny block i AM locky (‘L’ = Locked). Žádný z čítačů nehlásí chybu v žádném směru. Fyzická vrstva nehlásí žádné BIP chyby a mapování linek souhlasí správně se zapojením optických linek (Lane mapping = 0 1 2 3). Informace o rychlostech, které

udávají rychlost a verzi ethernetu 40G nejsou chybné. Je to způsobené převzetím Management modulu z 40G verze. Tento nedostatek se vyřeší úpravou adresového dekodéru na správné informace.

```
---------------------------- Ethernet interface 0 ----
Speed                        : 40 Gb/s
Transceiver status           : OK
Transceiver cage             : QSFP-0
Transceiver lane(s)          : 0|1|2|3
---------------------------------------- PMA regs ----
Link status                  : UP | UP
Speed                        : 40 Gb/s
Transmit Fault               : No
Receive Fault                : No
PMA type                     : 40GBASE-KR4
Supported PMA types ->
 *                           : 40GBASE-SR4
 *                           : 40GBASE-LR4
 *                           : PMA remote loopback
---------------------------------------- PCS regs ----
Link status                  : UP | UP
Speed                        : 40 Gb/s
Transmit Fault               : No
Receive Fault                : No
Global Block Lock            : Yes | No
Global High BER              : No | No
BER counter                  : 0
Errored blocks               : 0

Block status for lines       : L L L L
AM lock                      : L L L L
Lane mapping
  0  1  2  3
BIP error counter
     0     0     0     0
```

Obr. 6.2: Výstup z konzole serveru s připojenou akcelerační kartou

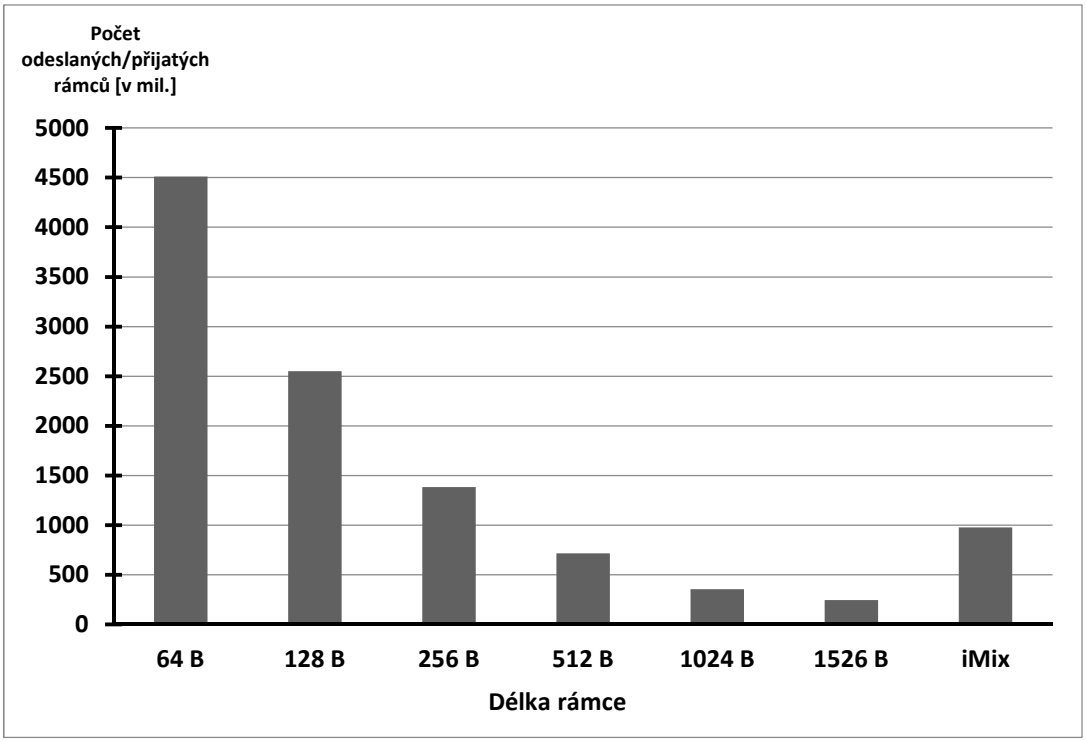
Hlavním testem implementace byl test chybovosti rámců v režimu zpětné vazby, kdy příchozí rámce ze Spirent modulu byly opět poslány zpět. Byly sledovány čítače příchozích a odchozích rámců, chybné rámce a čítače BIP (Bit Interleaved Parity) chyb, což jsou chyby paritních bitů. Hodnoty těchto čítačů byly sledovány jak na straně Spirent TestCenter, tak na výstupu z konzole ze serveru připojené karty. Test byl prováděn pro verianty délek rámců 64B, 128B, 256B, 512B, 1024B, 1526B (maximální možná délka ethernetového rámce pro fyzickou vrstvu) a režimem iMix, který simuluje reálný provoz (náhodné délky rámců s náhodnou četností). Každý provoz byl spuštěn po dobu ~60 s při zatížení linky 100%, tedy 50 Gb/s. Výsledky jednotlivých testů a stavy čítačů zobrazují Tab. 6.3 a Tab. 6.4 a graf na Obr. 6.3.

Tab. 6.3: Výsledky testů fyzické vrstvy pro různé délky rámců (Spirent TestCenter)

Délka rámce [B]	Stavy čítačů			
	Odeslané rámce	**Přijaté rámce**	**Chybné rámce**	**BIP chyby**
64	4 509 948 033	4 509 948 033	0	0
128	2 552 435 031	2 552 435 031	0	0
256	1 385 803 643	1 385 803 643	0	0
512	718 157 157	718 157 157	0	0
1024	357 327 729	357 327 729	0	0
1526	245 783 647	245 783 647	0	0
iMix	979 605 555	979 605 555	0	0

Tab. 6.4: Výsledky testů fyzické vrstvy pro různé délky rámců (výstup z konzole)

Délka rámce [B]	Stavy čítačů			
	Přijaté rámce	**Zpracované rámce**	**Chybné rámce**	**BIP chyby**
64	4 509 948 033	4 509 948 033	0	0
128	2 552 435 031	2 552 435 031	0	0
256	1 385 803 643	1 385 803 643	0	0
512	718 157 157	718 157 157	0	0
1024	357 327 729	357 327 729	0	0
1526	245 783 647	245 783 647	0	0
iMix	979 605 555	979 605 555	0	0



Obr. 6.3: Závislost odeslaných a přijatých ethernetových rámců za jednotku času na délce rámce

Z výsledků všech testů je zřejmé, že fyzická vrstva je plně funkční co se týče zpracování dat z příjmací RX a vysílací TX části. To potvrzuje také nulový počet chyb paritních bitů (BIP). Můžeme vidět, že jak čítače ze Spirent TestCenter, tak čítače z výstupu konzole jsou totožné, což eliminuje možnou chybu některého z výstupu.

Z Obr. 6.3 je vidět, že délka rámce odpovídá výslednému počtu zpracovaných rámců. Tedy zvýší-li se délka rámce dvakrát, sníží se zhruba dvakrát i počet zpracovaných rámců za stejnou jednotku času (~60 s).

7. ZÁVĚR

V rámci práce byl proveden rozbor problematiky fyzické vrstvy 50G ethernetu podle specifikací 25 Gigabit Ethernet Consortium. Byla rozebrána funkce jednotlivých podvrstev fyzické vrstvy s důrazem na vrstvy PCS a PMA. Rozbor zahrnoval také prostudování specifikací pro gigabitové transcievery pro různé architektury obvodů FPGA, a to konkrétně specifikace pro transciever GTY pro UltraScale+ a transciever GTZ pro Virtex 7.

Na základě tohoto rozboru byl vytvořen návrh fyzické vrstvy pro jednotlivé podvrstvy a implementované komponenty. Pomocí tohoto návrhu byl vytvořen popis fyzické vrstvy v jazyce VHDL, který pomocí generických parametrů umožňuje jednoduše volit mezi cílovou architekturou FPGA obvodu (UltraScale+ nebo Virtex 7) a počtem 50G portů (jednoportová nebo dvouportová verze). Testování výsledného popisu probíhalo na akcelerační kartě s architekturou FPGA obvodu UltraScale+ a dvěma optickými transcievery připojenými ke Spirent modulu. Při testování a ladění se narazilo na mnoho nečekaných potíží, jejichž řešení bylo velmi časově náročné, a to hlavně z důvodu dlouhého překladu samotného designu, který trval i několik hodin. Z důvodu časových problémů se v rámci práce stihla odladit pouze verze pro architekturu UltraScale+ s jednoportovou variantou. Avšak testování a ladění zbylých verzí, včetně dvouportové verze pro UltraScale+, probíhá i nadále v rámci zapojení do projektů firmy CESNET, z. s. p. o., z niž také zadání práce vzniklo.

Přínosem této práce je vytvořené řešení fyzické vrstvy 50G ethernetu pro obvody FPGA v jazyce VHDL, a připravený popis této vrstvy k využití v budoucích projektech firmy CESNET, z. s. p. o., které by tento režim ethernetu vyžadovali.

Literatura

[1] IEEE COMPUETER SOCIETY. *IEEE Standard for Ethernet*. New York: The Institute of Electrical and Electronics Engineers, 2016. ISBN 978-1-5044-0078-7

[2] SPURGEON, Charles a Joann ZIMMERMAN. *Ethernet: The Definitive Guide*. 2nd Edition. Sebastopol: O'Reilly Media, 2014. ISBN 978-1-449-36184-6.

[3] THE 25 GIGABIT ETHERNET CONSORTIUM. *25G & 50G Specification* [online]. [cit. 12. 11. 2018]. Dostupné z: https://25gethernet.org/

[4] Netcope Technologies. *Netcope FPGA Boards* [online]. [cit. 7. 11. 2018]. Dostupné z: https://www.netcope.com/en/products/fpga-boards

[5] Xilinx: *UltraScale Architecture GTY Transceivers: User Guide* [online]. 23. 6. 2014, poslední aktualizace 20. 9. 2017 [cit. 21. 10. 2018]. Dostupné z: https://www.xilinx.com/support/documentation/user_guides/ug578-ultrascale-gty-transceivers.pdf

[6] Programovatelná logika II: FPGA. In: *Abc Linuxu* [online]. 16. 1. 2013, poslední aktualizace 17. 1. 2013 [cit. 25. 10. 2018]. Dostupné z: http://www.abclinuxu.cz/blog/digital_design/2013/1/programovatelna-logika-ii-fpga

[7] PINKER, Jiří a Martin POUPA. *Číslicové systémy a jazyk VHDL*. 1. vyd. Praha, ČR: BEN - technická literatura, 2006, 349 s. ISBN 80-730-0198-5.

[8] Xilinx: *7 Series FPGAs GTZ Transceivers: User Guide*. 3. 4. 2012, poslední aktualizace 6. 1. 2015 [cit. 21. 10. 2018]. Proprietární dokument

[9] Xilinx: *UltraScale Architecture Libraries Guide* [online]. 19. 11. 2014 [cit. 21. 2. 2019]. Dostupné z: https://www.xilinx.com/support/documentation/sw_manuals/xilinx2014_4/ug974-vivado-ultrascale-libraries.pdf

[10] Xilinx: *7 Series FPGAs Clocking Resources: User Guide* [online]. 1. 3. 2011, poslední aktualizace 30. 7. 2018 [cit. 21. 3. 2019]. Dostupné z: https://www.xilinx.com/support/documentation/user_guides/ug472_7Series_Clocking.pdf

Seznam symbolů, veličin a zkratek

LUT	Look-Up Table
CLB	Configurable Logic Block
PMA	Physical Medium Attachment
PCS	Physical Coding Sublayer
PMD	Physical Medium Dependent
FEC	Forward Error Correction
MGT	Multi-Gigabit Transciever
GT	Gigabit Transciever
FPGA	Field-programmable Gate Array
VHDL	VHSIC Hardware Description Language
VHSIC	Very High Speed Integrated Circuit
PLD	Programmable Logic Device
RAM	Random Acces Memory
ROM	Read Only Memory
DSP	Digital Signal Processor
IOB	Input-Output Block
PSM	Programmable Switch Matrix
MII	Media Independent Interface
RX	Receive
TX	Transmit
PLL	Phase-Locked Loop
ILA	Integrated Logic Analyzer
JTAG	Joint Test Action Group
IP	Intellectual Property