

AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Replikacja danych

Opracowanie: Sylwia Miedzińska

Zespół w składzie Sylwia Miedzińska Piotr Michno Michał Mielec Konrad Łata Paweł Kozub

Wydział Inżynierii Metali i Informatyki Przemysłowej

Kraków, 09.06.2016



Zagadnienie replikacji [3][4]

 Replikacja danych polega na przechowywaniu kopii danych na wielu komputerach.

Podstawowy model zarządzania replikowanymi danymi:

- Klienci wysyła żądanie do systemu(front-end).
- System(front-end) zapewnia przejrzystość, ukrywając fakt, że dane są replikowane.
- System(front-end) komunikuje się z jednym lub więcej menedżerów repliki do pobierania / zapisywania danych.
- Zarządcy repliką współdziałają w celu zapewnienia, że dane są spójne.



Powody zwielokrotniania [1]

- Skalowalność możliwe jest rozłożenie obciążenia pomiędzy wieloma serwerami. Operacje zapisu i aktualizacji rekordów odbywają się na jednym serwerze, a pobieranie i przeszukiwanie danych z drugiego.
- Bezpieczeństwo Zwielokrotnianie w celach efektywnościowych jest ważne w sytuacjach, w których system rozproszony należy skalować w wymiarze liczbowym i geograficznym. Skalowanie w wymiarze liczbowym występuje na przykład wówczas, gdy wzrasta liczba procesów wymagających dostępu do danych zarządzanych przez jeden serwer. W tym wypadku efektywność możemy polepszyć przez zwielokrotnienie serwera i podział pracy



Podział wybranych algorytmów na pasywne i aktywne

Algorytmy pasywne	Algorytmy aktywne
Kopia podstawowa – pisanie/czytanie tylko z podstawowej	Brak kopii podstawowej - z globalnym znacznikiem czasu Lamporta
Kopia podstawowa – zapasowa – czytanie z kopii zapasowej	Brak kopii podstawowej – z procesem porządkowym (koordynatorem nadającym unikalne id operacjom)
	Brak kopii podstawowej – z głosowaniem kworum

Tabela 1. Podział wybranych algorytmów na pasywne i aktywne [10]



Nazwa algorytmu	Nazwisko Autora	Data powstania
Kopia podstawowa – pisanie/czytanie tylko z podstawowej	Budhijara	1993
Kopia podstawowa – zapasowa – czytanie z kopii zapasowej	Li, Hudak	1989
Brak kopii podstawowej - z globalnym znacznikiem czasu Lamporta	Rodrigues	1996
Brak kopii podstawowej – z procesem porządkowym (koordynatorem nadającym unikalne id operacjom)	Fonseca H	1996
Brak kopii podstawowej – z głosowaniem kworum	Thomas, Gifford	1979

Tabela 2. Zestawienie twórców algorytmów i roku powstania.[1]



OPIS ALGORYTMÓW



Kopia podstawowa – pisanie/czytanie tylko z podstawowej[4]

Wady i problemy związane z użyciem Nazwa algorytmu Opis algorytmu algorytmu Kopia podstawowa -Operacje pisania i czytania - Potencjalnym problemem efektywności w tym schemacie może być dość długi czas, który mija, pisanie/czytanie tylko z oparte sa na jednym podstawowej zanim procesowi inicjującemu aktualizację zdalnym serwerze Rezultat: dane nie sa zezwoli się na dalszą pracę. Wskutek tego zwielokrotniane lecz wszystkie aktualizacje realizujemy jako operacje umieszczone na jednym blokowane. Możemy też zastosować metodę bez blokowania. Gdy tylko serwer główny uaktualni zdalnym serwerze. swoją lokalną kopię x, zwraca potwierdzenie. Proces, które chce wykonać Dopiero potem powiadamia serwery zapasowe, operacje zapisania jednostki aby też wykonały uaktualnienia. danych x, przekazuje tę - Główny problem w nieblokowanych protokołach operację do serwera podstawa-zapas dotyczy tolerowania awarii. W ałównego x. Serwer ten schemacie z blokowaniem procesu klient wie na wykonuje uaktualnienie na pewno, że aktualizacja została wykonana na kilku lokalnej kopii x, po czym innych serwerach zapasowych. Pewności tej nie przekazuje uaktualnienie do ma w rozwiązaniu bez blokowania serwerów zapasowych. Każdy serwer zapasowy dokonuje również Klient aktualizacji i wysyła Jeden serwer potwierdzenie z powrotem iednostki x Serwer zapasowy do serwera podstawowego. Gdy wszystkie serwery zapasowe uaktualnia swoje kopie lokalne, wówczas serwer podstawowy wysyła Pamięć danych potwierdzenie do procesu, który zapoczatkował te działania R₁ - zamówienie czytania W₁ – zamówienie pisania przekazanie zamówienia przekazanie zamówienia do serwera x do serwera x - wykonano potwierdzenie zapisu R₃ - zwrócenie odpowiedzi W₄ - wykonano potwierdzenie zapisu R₄ – zwrócenie odpowiedzi



Kopia podstawowa – zapasowa – czytanie z kopii zapasowej [5]

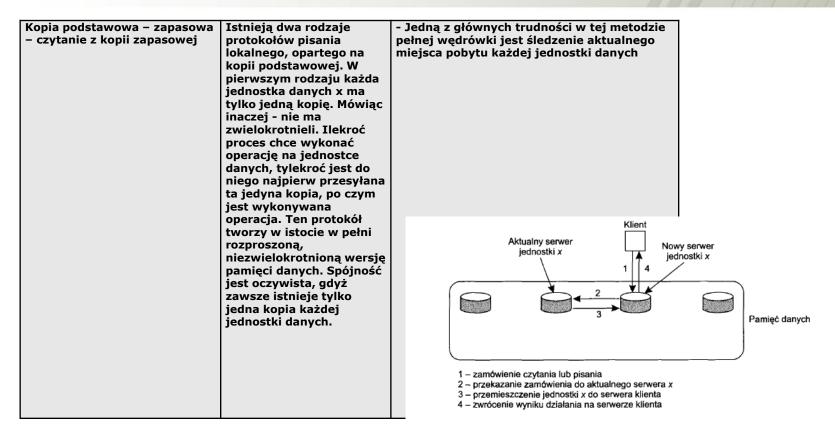


Tabela 3b Opis algorytmów i problemy związane z ich użyciem



Brak kopii podstawowej - z globalnym znacznikiem czasu Lamporta[6]

Brak kopii podstawowej - z globalnym	W protokołach zwielokrotnionych	- konieczność wykonywania operacji wszędzie w tym
znacznikiem czasu Lamporta	zapisów (ang. replicated-writes)	samym porządku
	operacje pisania możemy	- znaczniki czasu Lamparta źle się skalują w wielkich
	wykonywać na wielu kopiach, a	systemach rozproszonych
	nie tylko na jednej, jak w	
	wypadku kopii podstawowych.	
	Potencjalnym problemem	
	aktywnego zwielokrotnienia jest	
	konieczność wykonywania	
	operacji wszędzie w tym samym	
	porządku. Jest więc potrzebny	
	mechanizm całkowicie	
	uporządkowanego rozsyłania.	
	Rozsyłanie takie możemy	
	zrealizować przy użyciu	
	znaczników czasu Lamporta.	
	-	



Brak kopii podstawowej – z procesem porządkowym (koordynatorem nadającym unikalne id operacjom)[1]

Brak kopii podstawowej - z procesem porządkowym (koordynatorem nadającym unikalne id operacjom)	Każdą operacje przekazuje się najpierw porządkowemu Porządkowy przypisuje niepowtarzalny numer Porządkowy przekazuje tą operacje do wszystkich kopii. Uwaga! Operacie wykonywane są w kolejności numerów porządkowych.	- problem zwielokrotnionych wywołań
---	--	-------------------------------------

Tabela 3d Opis algorytmów i problemy związane z ich użyciem



Brak kopii podstawowej – z głosowaniem kworum[1]

Brak kopii podstawowej - z głosowaniem kworum

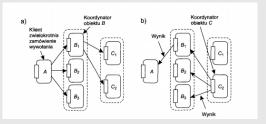
plików , zakładamy że plik jest zwielokrotniany na N serwerach W celu zaktualizowania pliku klient musi sie skontaktować przynajmniej z połowa + 1 serwerów.

Po zgodzie serwerów plik zostaje podmieniony a nowa wersja zostaje zaopatrzona nowym numerem – numer wersji służy do identyfikowania wersji pliku i jest taki sam dla wszystkich nowo zaktualizowanych kopii pliku. Aby przeczytać zwielokrotniony plik klient również musi się skontaktować z ponad połowa serwerów i prosić je o wysłanie numerów wersii.

Rozważamy rozproszony system - Uproszczony schemat Gifforda: Do czytania pliku mającego N zwielokrotnień wymaga się od klienta kworum czytania, Czyli dowolnego zbioru N. lub wiecej serwerów, podobnie do zmodyfikowania pliku potrzeba kworum pisania, co najmniej N_w serwerów, Wartości te musza spełniać następujące ograniczenia

$$N_{r+}N_w>N$$

 $N_w>N/2$



Na rysunku a może wystąpić konflikt pisanie – pisanie ponieważ nie iest spełniony warunek 1. W szczególności gdy jeden klient odbierze swój zbiór do zapisu {A, B, C, E, F, G}, a drugi wybierze {D, H, I, J, K, L}

Na rysunku b jest szczególnie interesująca sytuacja ponieważ N, wynosi tu 1 co oznacza możliwość czytana zwielokrotnionego pliku za pomocą dowolnej liczby kopii – co oznacza aktualizowanie wszystkich kopi przy zapisywaniu – nazwa schematu – czytaj jedno zapisuj wszystko.



Zakres stosowalności

Algorytmy, w których każda kopia umożliwia czytanie są wykorzystywane w bardzo rozproszonych serwerach, np. DNS lub CDN.

Aktywa replikacja danych znajduje zastosowanie w rozproszonych systemach plików.(OpenAFS, Google Cloud Starage)

Protokół Gossip, system BAYOU, *Coda* (Constant Data Availability)[2]



Wybór algorytmu

- Wybrany algorytm jest najprostszy i najłatwiejszy w implementacji
- Wybrany algorytm nie zawiera w sobie dodatkowych zagadnień z zakresu systemów rozproszonych.
- Wybrany algorytm pozwala w najbardziej przejrzysty i czytelny sposób przestawić zasadę działania algorytmów replikacji.



Wybór technologii

AGH					
SOAP[11]	RMI[12]	Sockety[13]	REST[14]	MPI[18]	CORBA[15]
 Simple Object Access Protocol, protokół komunikacyjn y oparty o XML. Niezwykłą elastyczność protokołu, który pozwala przenosić właściwie dowolne informacje Duży narzut samego języka XML (rozmiar komunikatu jest znacząco większy niż sumaryczny rozmiar danych w nim zawartych) 	Remote Method Invocation, umożliwia programowani e rozproszone w Javie. Mechanizm zdalnych wywołań umożliwia wywołanie metod z obiektów pod kontrolą innych maszyn wirtualnych języka Java. Mogą działać na różnych komputerach.	Narzędzie do komunikacji pomiędzy procesem działającym na tej samej maszynie bądź na innym. Do stworzenia socketa potrzeba: protokołu, domeny oraz typu komunikacji.	REpresentional State Transfer, zamiast XML używa prostego URL. Większość zadań można uzyskać poprzez żądania HTTP 1.1 takie jak GET, POST, PUT, DELETE. Dane można przesyłać przez JSON, RSS.	Message Passing Interface, protokół przesyłania komunikatów pomiędzy procesami programów równoległych. Komunikacja może być grupowa bądź punktowa. MPI_Init – inicjalizacja MPI, MPI_Send – wysyłanie blokujące, MPI_Recv – odbiór blokujący MPI_Finalize – zakończenie działania	Common Object Request Broker Architecture, przeznaczona przede wszystkim do wspomagania programowania pomiędzy systemami niekompatybiln ymi. Określa metody dostępu do obiektów i komunikacji między obiektami

Tabela 4 Zestawienie wybranych technologii



Diagramy UML

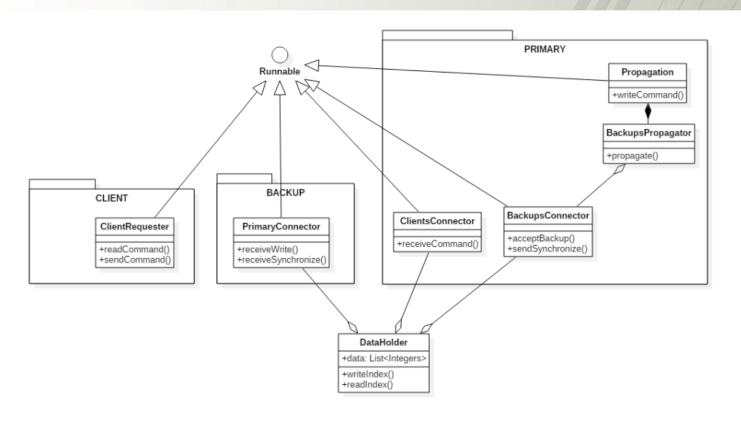
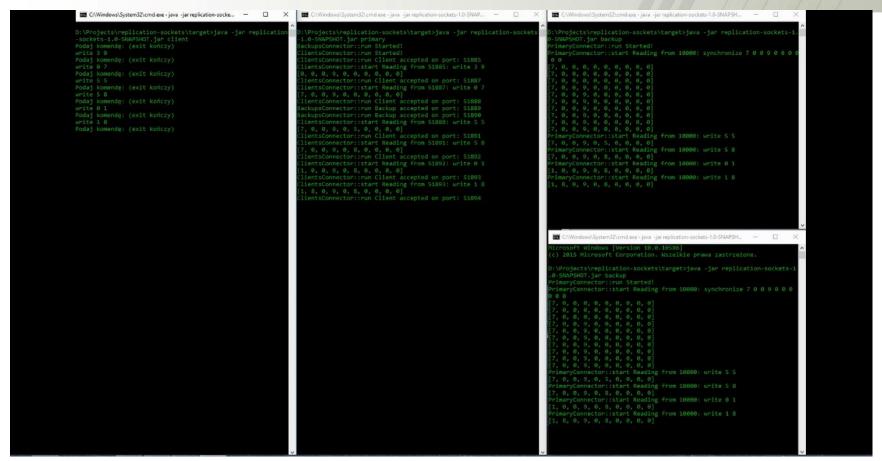


Diagram 1. Diagram UML Kopia podstawowa – pisanie zdalne (technologia sockety)



Działanie algorytmu w technologii socket



Screen z działania w technologii Socket



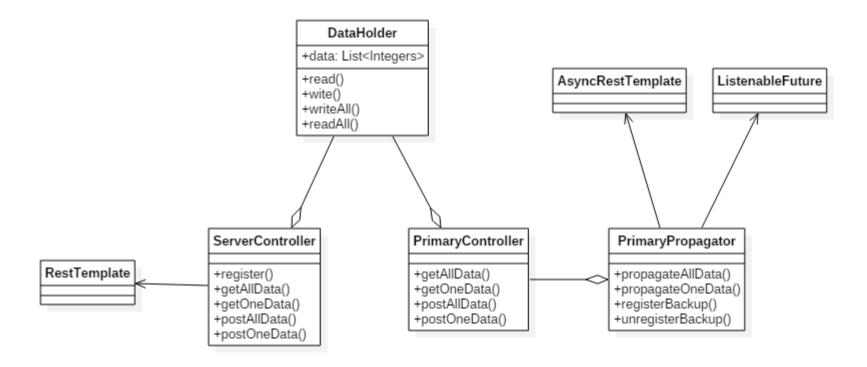
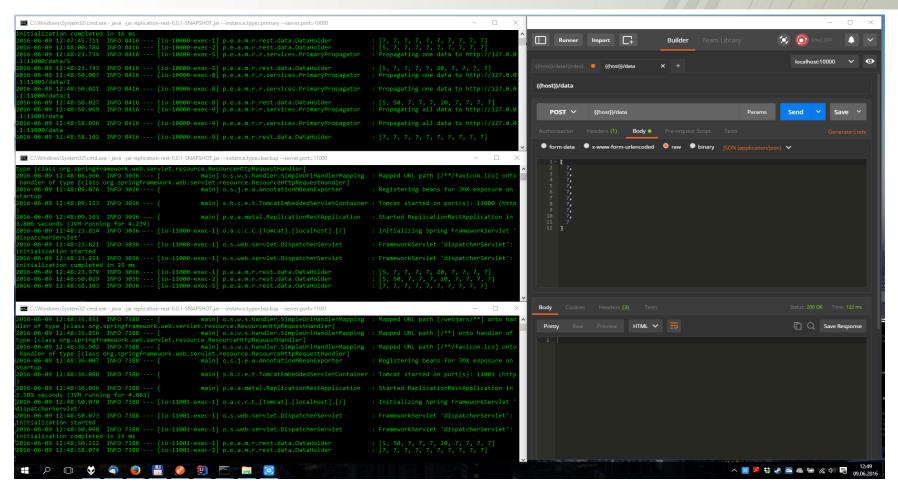


Diagram 2. Diagram UML Kopia podstawowa – pisanie zdalne (technologia REST)



Działanie algorytmu w technologii rest





Analiza porównawcza implementacji

Czasowy nakład pracy

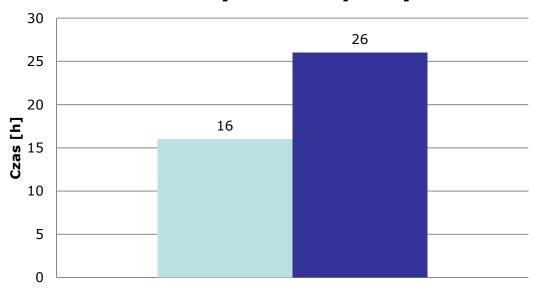


Tabela 5 Nakład czasu pracy w
poszczególnych technologiach

Czas [h]

16

26

Nakład pracy

Sockety

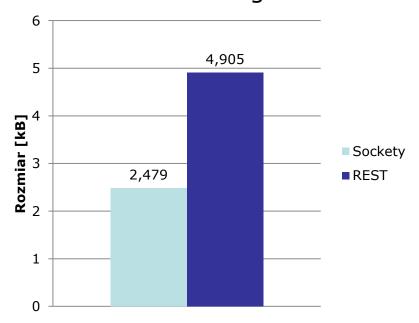
REST

Wykres 1 Nakład czasu pracy w poszczególnych technologiach



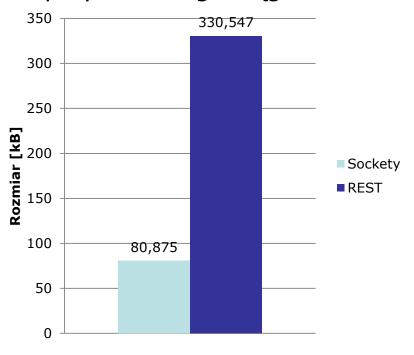
porównania szybkości działania tego samego algorytmu w różnych technologiach

Wielkość wiadomości dla przykładowego łańcucha znakowego



Wykres 2 Wielkość wiadomości dla przykładowego łańcucha znakowego

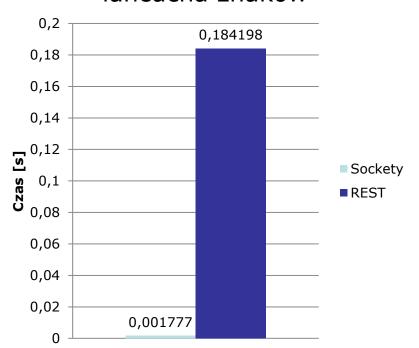
Wielkości wiadomości dla przykładowego ciągu liczb



Wykres 3 Wielkości wiadomości dla przykładowego ciągu liczb

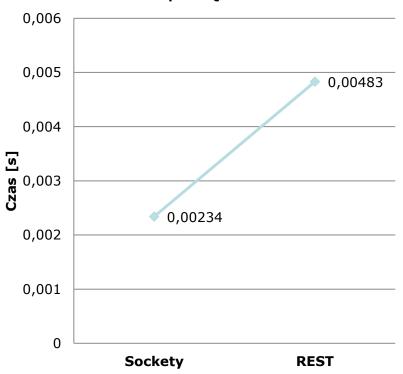


Czas przesyłania dużego łańcucha znaków



Wykres 5 Czas przesyłania dużego łańcucha znaków

Czas połączenia



Wykres 4 Czas połączenia



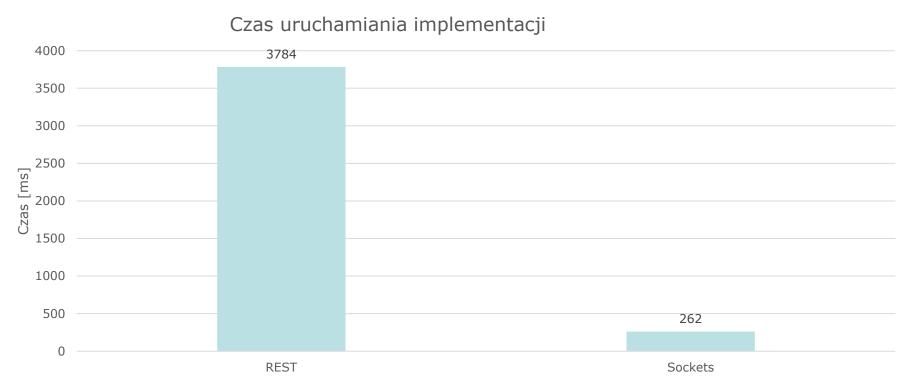
Wykres ilości kodu



Wykres 6 llości kodu w poszczególnych technologiach.



Czas uruchamiania implementacji



Wykres 7 Czas uruchamiania implementacji.



Źródła

- 1. Andrew S.Tanenbaum, Maarten Van Steen, "Distribiuted Systems Principles and Paradigms" Second Edition, 2007, ISBN 0-13-239227-5
- 2. Sukumar Ghosh, "Distributed Systems An Algorithmic Approach", 2007, ISBN 1-58488-564-5
- 3. George Coulouris, "Distributed Systems Concepts and Design" Fifth Edition, 2012, ISBN 0-13-214301-1
- 4. Mullender Sape, "Dustributed Systems" Second Edition, 1993, ISBN 978-0201624275
- 5. Li K., Hudak P., "Memory Coherence in Shared Virtual Memory Systems", 1989
- 6. Fonseca H., Verissimo P., "Totally Ordered Multicast in Large-Scale Systems", Opublikowano: "Proceedings of the 16th International conference on Distributed Computing Systems", 1996, ISBN 0-8186-7399-0
- 7. Stanford University, "Chapter 14 Replication", http://www-cs-students.stanford.edu/~dbfaria/quals/summaries/Coulouris-chap14.txt (dostęp 6.05.2016)
- 8. Sami Rollins, "Replication", 10.2008 http://www.cs.usfca.edu/~srollins/courses/cs682-s08/web/notes/replication.html (dostęp 10.05.2016)
- 9. K. Banas "Systemy Równoległe i Rozproszone Wykład 13", 03.2016 http://www.metal.agh.edu.pl/~banas/SRR/SRR_W13_Rozglaszanie_Uzgadnianie.pdf (dostęp 12.05.2016)



- 10. Politechnika Warszawska, "Rozproszone systemy operacyjne", 06.2007, http://www.ia.pw.edu.pl/~tkruk/edu/rsob2010/rso_proj2007/rso2007 (dostęp 1.06.2016)
- 11. Oracle, "Simple Object Access Protocol Overview", 2001, https://docs.oracle.com/cd/A97335_01/integrate.102/a90297/overview.htm (dostęp 12.05.2016)
- 12. Polsko-Japońska Akademia Technik Komputerowych, "RMI programowanie rozproszone", 2010, http://edu.pjwstk.edu.pl/wyklady/mpr/scb/W11/W11.html (dostęp 20.05.2016)
- 13. M.Zakrzewicz, "Wprowadzenie do technologii Web Services: SOAP, WSDL i UDDI", 05.2006, http://www.cs.put.poznan.pl/mzakrzewicz/pubs/ploug06ws.pdf (dostęp 20.05.2016)
- 14. J.Brzeziński, C.Sobaniec Politechnika Poznańska "Usługi sieciowe REST", 2013 https://www.soa.edu.pl/c/document_library/get_file?uuid=46b0faf6-6743-4184-ab16-dbddfd413685&groupId=10122 (dostęp 20.05.2016)
- 15. T.Olas Politechnika Częstochowska, 2011 "Oprogramowanie systemów równoległych i rozproszonych", http://icis.pcz.pl/~olas/srr/wyklad8.4.pdf (dostęp 2.06.2016)
- 16. Robert Werembel "Rozproszeone bazy danych replikacja danych (Wykład 1) http://wazniak.mimuw.edu.pl/images/5/55/ZSBD-2st-1.2-lab1.tresc-1.1.ppt(dostęp 8.06.2016)
- 17. Ph. D. Simon Tuffs, "How fast is your network today?", 2004 http://soap-stone.sourceforge.net/ (dostep 08.06.2016)
- 18. K.Banaś "Programowanie równoległe wykład 10" 2015, http://www.metal.agh.edu.pl/~banas/PR/PR_W10_MPI_wstep.pdf (dostęp 08.06.2016)
- 19. University of California, "Architectural Styles and the Design of Network-based Software Architectures", 2000, https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm (dostęp 03.06.2016)
- 20. Ph. D. Martin Fowler, "Rishardson Maturity Model", 2010, http://martinfowler.com/articles/richardsonMaturityModel.html (dostęp 02.06.2016)
- 21.Oracle Java Documentation "All About Sockets", 2015, https://docs.oracle.com/javase/tutorial/networking/sockets/ (dostęp 04.06.2016)